UNIVERSIDADE FEDERAL DO PARANÁ

RÉRIS APARECIDA PEREIRA DE LIMA

MINERAÇÃO DE DADOS ABERTOS COM A FERRAMENTA *WEKA*: MICRODADOS CAGED DE PESSOAS COM DEFICIÊNCIA DA REGIÃO SUL DO BRASIL

CURITIBA

2017

RÉRIS APARECIDA PEREIRA DE LIMA

MINERAÇÃO DE DADOS ABERTOS COM A FERRAMENTA *WEKA*: MICRODADOS CAGED DE PESSOAS COM DEFICIÊNCIA DA REGIÃO SUL DO BRASIL

Trabalho de Conclusão de Curso apresentado como requisito parcial à obtenção de grau de Bacharel do Curso de Gestão da Informação do Setor de Ciências Sociais Aplicadas da Universidade Federal do Paraná.

Orientador: Prof. Dr. Celso Yoshikazu Ishida

CURITIBA

2017

À minha mãe Davina Pereira (*in memoriam*), que que se faz presente em todos os dias da minha vida, sei que de seu lugar olha por mim, sofre com minhas derrotas e rejubila comigo em minhas vitórias.

Também dedico este trabalho ao meu pai, ao meu namorado e às minhas irmãs que, com muito carinho e apoio, não mediram esforços para que eu chegasse até esta etapa de minha vida.

AGRADECIMENTOS

Em primeiro lugar à Deus, pela força e coragem durante toda esta longa caminhada. Sem Ele, nada disso seria possível.

Ao meu pai João de Jesus Martins de Lima e à minha irmã Selma Pereira da Mota, por tudo que fizeram por mim ao longo de minha vida. Desejo poder ter sido merecedora do esforço dedicado por vocês em todos os aspectos, especialmente quanto à minha formação.

Com muito amor, uma eterna gratidão ao meu namorado, Ademar da Silva Junior - por sua compreensão, carinho, presença e incansável apoio ao longo do período da graduação, principalmente da elaboração deste trabalho.

Às minhas irmãs Angela Pereira de Lima e Suelen Pereira de Lima pelo apoio que transcende o tempo e a distância, vocês são minha motivação diária e inspiração de vida.

Ao meu mestre e orientador Celso Yoshikazu Ishida, pela grande sabedoria, paciência, incentivo, orientação e principalmente por acreditar em mim, possibilitando concluir mais esta etapa da minha vida acadêmica.

À Elizabeth Senna da Silva, Luci Mara Hervis e Chao Chung Fan: vocês se tornaram um exemplo de vida, de luta diária e de fé. Obrigada pelos valores que me ensinaram e pelo apoio diário, que foram fundamentais para que eu chegasse até aqui.

À todos os professores do curso, que foram tão importantes na minha vida acadêmica: agradeço pelo conhecimento compartilhado, pelo convício e pelo incentivo profissional.

Ao meu querido amigo Iago Lopes e minhas amigas especiais que não posso deixar de citá-las aqui: Caroline Dantas, Barbara Silva, Talita Rampão, Jeslin Elen, Larissa Liu, Franciele Seixas, Amanda Cruz, Mayara Pinto e Auélica Souza: Obrigada por acreditarem em mim, pelas alegrias e ansiedades compartilhadas, pela força, companheirismo e amizade.

Há muitas outras pessoas especiais para agradecer, as quais simbolizo na minha tia do coração Beatriz Dantas Campos e irmãs Elizangela Pereira Mota e Regina Pereira Mota: Obrigada pelas orações e pela força que transcende esta distância todos os dias e principalmente por acreditarem em mim.

À equipe do departamento de Pós-Graduação da Universidade Positivo, pela confiança, compreensão e apoio constante enquanto estive com vocês.

À todos aqueles que de alguma forma estiveram próximos a mim durante esta longa jornada e ainda estão, fazendo esta vida valer cada vez mais a pena.

“A maior recompensa para o trabalho do homem não é o que ele ganha com isso, mas o que ele se torna com isso”.

(John Ruskin)

RESUMO

O presente trabalho refere-se à aplicação de métodos de mineração de dados, do tipo classificação, na base de dados abertos CAGED (Cadastro Geral de Empregados e Desempregados), especialmente no que tange aos dados de mercado de trabalho das pessoas com deficiência (PcD) da região Sul do Brasil, pois, atualmente o CAGED é uma das importantes fontes de informação do mercado de trabalho de âmbito nacional e de periodicidade mensal. De natureza aplicada, abordagem quantitativa, com delineamento bibliográfico e de caráter exploratória, tendo como metodologia prática as etapas do KDD (*Knowledge Discovery in Databases*). Os testes práticos definiram que para a tarefa de classificação, os métodos com melhores desempenhos são o J48 e PART, das heurísticas *tree* e *rules*, respectivamente. Não descobriu-se padrões de comportamento válidos sobre o mercado formal de trabalho das pessoas com deficiência, visto que os resultados obtidos foram complexos e estatisticamente mostraram-se insatisfatórios, porém, foi possível identificar os principais atributos que se constituem como um dos importantes fatores para a determinação dos objetivos pretendidos (padrão de desligamento/admissão e perfil das PcD na referida região). Também foi possível obter conhecimento válido e descrever o mercado de trabalho dos PcD na referida região com base na estatística descritiva, como por exemplo, que os deficientes físicos e auditivos são os mais responsáveis pela movimentação de desligamento/admissão no presente mercado, que o saldo de movimentação do Sul seguiu a mesma tendência do Brasil dentro do período analisado, evidenciando que se o cenário econômico brasileiro está favorável, consequentemente isso refletirá no Sul com a geração de mais empregos e que admissão por reemprego e desligamento por demissão sem justa causa são os fatores determinantes que predominam nos registros de admissões e desligamentos, respectivamente.

Palavras-chaves: CAGED. Mineração de Dados. Dados Abertos. Pessoa com Deficiência. Mercado de Trabalho.

ABSTRACT

The current work refers to the application of data mining methods, of the classification type, in the open database CAGED (General Register of Employed and Unemployed), especially with regard to the labor market data of people with disabilities of the southern region of Brazil, because currently the CAGED is one of the important sources of information on the national and monthly labor market. Applied nature, a quantitative approach, with a bibliographic deliniation and exploratory nature, with KDD (Knowledge Discovery in Databases) as a practical methodology. Practical tests defined that the best performing methods are J48 and PART, of the tree and rule heuristics respectively, for the classification task. No valid patterns of behavior were found on the formal labor market of people with disabilities, in view of the results were complex and statistically unsatisfactory, but it was possible to identify the main attributes that constitute one of the important factors to determine the intended objectives (dismissal/admission pattern and people with disabilities profile in said region). Also it was possible to obtain valid knowledge and to describe the labor market of people with disabilities in that region based on descriptive statistics, for example, person with hearing loss or handicapped are the most responsible for the movimentation of the dismissal/admission in the said market, and the movimentation balance of the Southern region followed the same trend of the rest of Brazil within the analyzed period, evidencing that if the Brazilian economic scenario is therefore, this will reflect in the Southern region with the generation of more jobs and that admission for reemployment and dismissal without just cause are the determining factors that prevails in the admission and dismissal records, respectively.

Keywords: CAGED. Data Minning. Open Data. People with Disabilities. Labor Market.

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE SIGLAS E ABREVIATURAS

CAGED	Cadastro Geral de Empregados e Desempregados
CLT	Consolidação das Leis do Trabalho
DCBD	Descoberta de Conhecimento em Bases de Dados
ELKI	*Environment for Developing KDD-Applications Supported by Index-Structures*
FP RATE	*False Positive Rate*
GI	Gestão da Informação
IBGE	Instituto Brasileiro de Geografia e Estatística
INDA	Infraestrutura Nacional de Dados Abertos
KDD	*Knowledge Discovery in Databases*
LAI	Lei de Acesso à Informação
MTPS	Ministério do Trabalho e Previdência Social
PBDA	Portal Brasileiro de Dados Abertos
PcD	Pessoa com Deficiência
PMO	Ministério do Planejamento, Orçamento
SCIELO	*Scientific Eletrocnic Libary Online*
SPELL	*Scientific Periodicals Electronic Library*
STI	Secretaria de Tecnologia da Informação
TP RATE	*True Positive Rate*
WEKA	*Waikato Environment for Knowledge Analysis*

SUMÁRIO

1 INTRODUÇÃO

Na presente seção, contextualizou-se o problema de pesquisa, a justificativa do tema, os objetivos gerais e específicos, a metodologia de desenvolvimento e a estrutura de organização do documento.

1.1 CONTEXTUALIZAÇÃO DO PROBLEMA

O contexto atual de *Big Data* - termo que descreve o imenso volume de dados, é o resultado conjunto das áreas governamentais, acadêmicas e corporativas, os responsáveis pelo imenso volume de dados e informações, que só cresce por meio dos seus usos intensivos da tecnologia.

O Governo, em todas as suas esferas e órgãos responsáveis, “[...]é particularmente significativo a este respeito, tanto pela quantidade como pela centralidade dos dados recolhidos [...]”, de acordo com a *Open Knowledge Internacional* (2017, tradução nossa).

Ainda segundo a *Open Knowledge Internacional*, “[...] a maior parte dos dados do governo são dados públicos e, consequentemente, podem ser tornarem abertos e disponibilizados a outros utilizadores”, (2017, tradução nossa). Dados abertos são dados publicados no formato aberto e sob licença aberta, ou seja, dados que toda e qualquer pessoa pode acessar, manusear e compartilhar, independente da finalidade, desde que sua abertura e proveniência seja preservada.

O governo brasileiro particularmente, muito tem contribuído para a divulgação de seus dados: além da Lei de Acesso a Informação (LAI), Lei 12.527, sancionada no dia 11 de Novembro de 2011, que regula o acesso a dados e informações “detidas” pelo governo, também disponibiliza o Portal Brasileiro de Dados Abertos (PBDA), ferramenta que funciona como um catálogo, onde os dados publicados pelos órgãos do governo estão disponibilizados de uma forma organizada e simples, o que facilita a busca e a sua recuperação.

Um exemplo de dados abertos disponibilizados via o portal são os dados do CAGED - Cadastro Geral de Empregados e Desempregados, cujos microdados dizem respeito às admissões e demissões no mercado de trabalho formal brasileiro, podem também ser facilmente recuperados na página web do Ministério do Trabalho.

Na base de dados pública CAGED é possível distinguir os dados sobre o mercado formal de pessoas com deficiência (PcD), que segundo o art. 3 do Decreto 3.298/1999 são pessoas que possuem “Toda perda ou anormalidade de uma estrutura ou função psicológica, fisiológica ou anatômica que gere incapacidade para o desempenho de atividade, dentro do padrão considerado normal para o ser humano.”.

Segundo os dados coletados pelo Instituto Brasileiro de Geografia e Estatística - IBGE, no censo demográfico de 2010, as pessoas que possuem pelo menos um tipo de deficiência - visual, auditiva, motora e mental ou intelectual, representavam 23,09% da população total, ou seja, 45.606.048 milhões de pessoas. Segundo o IBGE, apenas 20,4 milhões com deficiências estavam ocupadas, representando 23,6% do total de 86,4 milhões de pessoas de 10 anos ou mais ocupadas.

Inúmeras iniciativas tentam garantir o direito do deficiente, tal como a popular Lei de cotas (Decreto 3.298/1999), onde no seu art. 36 estabelece que a empresa com 100 ou mais funcionários é obrigada a preencher de dois a cinco por cento dos seus cargos com pessoas com deficiência e reabilitadas, na seguinte proporção do número total de funcionários: até 200, 2%; de 201 a 500, 3%; de 501 a 1.000, 4%; de 1001 e acima, 5%.

Na atual Sociedade da Informação, as bases de dados de órgãos públicos constituem um recurso essencial, umas das principais fontes de informação, porém, as organizações, sejam elas públicas ou privadas, se mostram ineficientes quando se trata de aproveitá-los, seja para tomar decisões ou gerar conhecimento. Tal ineficiência se dá devido a limitação da capacidade humana de interpretação e suas habilidades técnicas, o que gera “uma necessidade urgente de novas [...] teorias computacionais e ferramentas para ajudar os humanos a extrair informações úteis (conhecimento) a partir do rápido crescimento dos volumes de dados digitais”, de acordo com Fayyad et al. (1996, p. 37)

Surge então a Mineração de Dados, diante da disponibilidade de um volume imenso de dados e da necessidade de transformá-los em informação útil e conhecimento para suporte à decisão. Devido sua característica multidisciplinar, não há um consenso na literatura sobre sua definição, o que vai variar segundo a área de atuação do respectivo autor.

Do inglês *Data mining*, a mineração de dados consiste no “processo de análise de conjuntos de dados que tem por objetivo a descoberta de padrões interessantes”

a fim de obter conhecimento, segundo Amorim (p. 11, 2006). Já Fayyad et al, complementam que:

> A mineração de dados é um passo no processo de Descoberta de Conhecimento que consiste na realização da análise dos dados e na aplicação de algoritmos de descoberta que, sob certas limitações computacionais, produzem um conjunto de padrões de certos dados. (FAYYAD et al, 1996, p.41, tradução nossa)

Tendo em vista as breves apresentações realizadas, a presente pesquisa visa a mineração dos dados pertinentes às pessoas com deficiência (PcD) da região Sul do Brasil provenientes da base CAGED a fim de responder as seguintes questões:

- Será possível identificar algum padrão de desligamento/admissão no mercado formal de trabalho das pessoas com deficiência (PcD) da região Sul do Brasil com aplicação de técnicas de mineração de dados sobre a base CAGED?
- Qual o perfil dos das PcD que movimentam o mercado de trabalho na região Sul?

Diante disso, foram analisadas 14 variáveis, entre elas o sexo, tipo deficiência, tipo do movimento desagregado, salário mensal, UF, etc. - extraídos da base de dados abertos CAGED (que traz informações sobre a admissão e demissão de todos empregos com carteira assinada).

1.2 JUSTIFICATIVA

Atualmente os órgãos públicos constituem umas das principais fontes de informação, pois, geram e armazenam dados dos diversos setores da sociedade e por meio da internet e revolução das tecnologias da informação e da comunicação (TIC), é possível o arquivamento de tais dados de forma digital e a sua disponibilização aos cidadãos.

A base de dados pública CAGED é uma das principais fontes de informação a todos os setores da sociedade brasileira quando se trata do mercado de trabalho formal brasileiro e no mesmo, é possível a distinção entre pessoas com deficiência e sem deficiência.

Diante do paradigma do mercado de trabalho para as pessoas com deficiência e o seu contexto histórico, que perdura até os dias atuais, onde sua participação é baixa quando comparado com as pessoas sem deficiência, a motivação do presente trabalho é a transformação de dados públicos (no caso, a base CAGED) em conhecimento de valor social com o uso da Descoberta De Conhecimento em Bases de Dados (DCBD ou *knowledge discovery in databases — KDD*) no que tange aos dados relacionados às pessoas com deficiência.

Outras motivações para tal objetivo é o interesse na área da pesquisa (Mineração de Dados) e o baixo volume de publicações da mesma quando se trata de dados abertos, principalmente no idioma português - tal conclusão se deve à análise bibliométrica (apresentada detalhadamente na subseção a seguir) no resultado do levantamento realizado no dia 09 de Maio de 2017 nas bases de dados *Scopus* e *Web of Science - Coleção Principal (WoS).*

1.2.1 Análise bibliométrica

A fim de justificar uma das motivações apresentadas na justificativa, que é o baixo volume de publicações que relacionam mineração de dados e dados abertos no idioma português, no dia 09 de Maio de 2017 realizou-se um levantamento nas principais bases de dados públicas e restritas, listadas a seguir: *Scientific Electronic Library Online (Scielo), Scopus, Scientific Periodicals Electronic Library (Spell) e Web of Science - Coleção Principal (WoS)*.

A pesquisa em todas as bases de dados foi de caráter avançado, combinando as seguintes palavras pelo operador de busca *E*, indicando que as duas condições devem ser atendidas no critério de busca:

- “Mineração de dados” E “dados abertos”;
- “*Data mining”* E “*Open data”*.

Nas bases de dados *Scielo* e *Spell* optou-se pelos campo de busca “Todos os índices” e “Palavra-chave” respectivamente, porém, não houve resultados para nenhuma das duas formas pesquisadas. Nas outras bases de dados, tais como *Scopus* e *WoS*, não houve resultados quando realizada a busca com os termos em português, o que corrobora com a justificativa.

Os resultados foram mais significativos quando combinado os termos em inglês - *Data mining* E *Open data* nas bases de dados *Scopus* e *WoS*. O resultado é conferido no Quadro 1 a seguir, apresentados segundo os respectivos campos de busca:

QUADRO 1 - RESULTADO DO LEVANTAMENTO DAS PALAVRAS-CHAVE RELACIONADAS NAS BASE DE DADOS RESTRITAS

Termos de busca / Base de dados	*Data mining* E *Open data*	Critério/Campo de busca	Data da pesquisa
Scopus	259	*Article title, Abstract, Keywords*	09/05/2017
Web of Science	70	Tópicos	09/05/2017

FONTE: A autora (2017).

A diferença é significativa entre ambas as bases porque a *Scopus* possui mais de 21 mil periódicos, de 5 mil editores internacionais, 24 milhões de patentes, além de outros documentos, enquanto a *WoS* possui pouco mais de 9.000 periódicos, segundo a CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior), ou seja, a *Scopus* possui uma cobertura maior, indexando mais que o dobro que a *WoS* indexa.

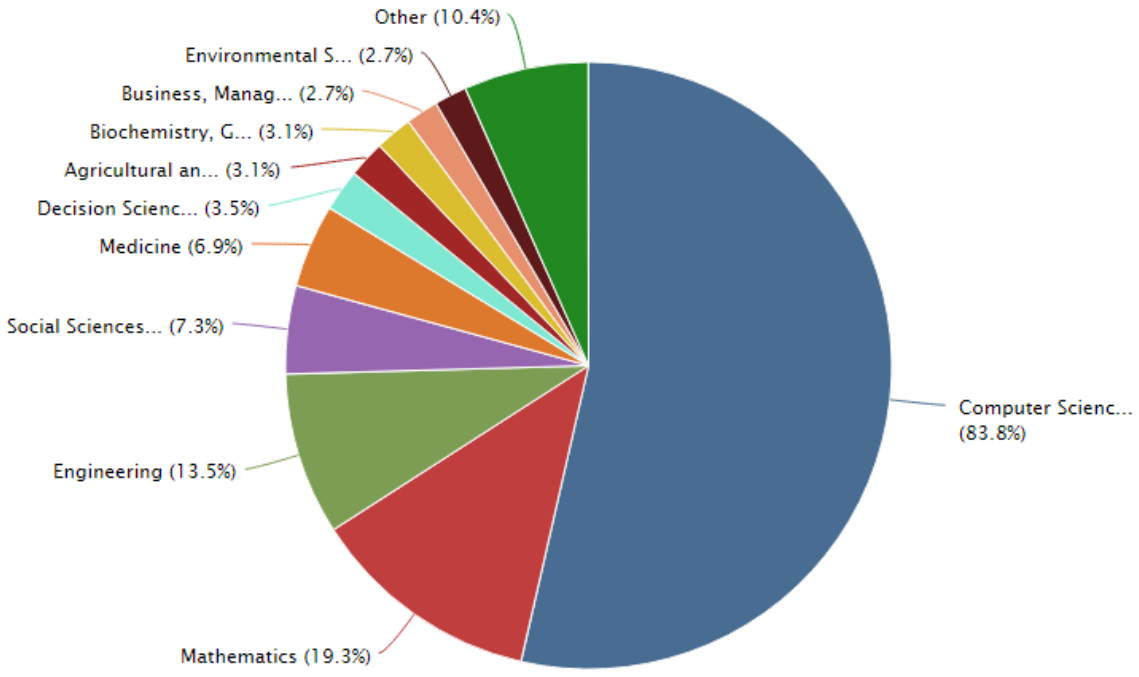
Segundo Yoshida (2010, p.58), a análise bibliométrica pode ser entendida como “uma metodologia de contagem sobre conteúdos bibliográficos, na sua essência.”

Com base nos resultados significativos apresentados no Quadro 1, aplicou-se técnicas bibliométrica para identificar os idiomas e as áreas de desenvolvimento de pesquisa mais predominantes sobres os respectivos resultados. Para isso, os resultados das duas bases de dados em destaque foram exportados (arquivos em

CSV - Documento separados por vírgula para os resultados providos da *Scopus* e arquivos em TXT - Documentos de texto, para os resultados providos da *Web of Science*) e abertos na ferramenta MS Excel para tal aplicação. Os resultados são apresentadas a seguir, por base de dados:

SCOPUS - A pesquisa com base na combinação das palavras "*Data mining*" e "*Open data*" na base de dados *Scopus* obteve 259 resultados. Constatou-se que o idioma inglês é predominante, representando 95,75% do resultado da respectiva busca (equivalente a 248 de 258 resultados). Em segundo lugar fica o chinês e o espanhol - ambos com 0,77% (equivalente a 2 resultados), o idioma português obteve apenas um resultado, correspondendo a 0,39% do total, juntamente com o francês, alemão, russo e outros registros sem dados. A *Scopus* permite analisar os resultados diretamente em sua página web conforme os parâmetros selecionados, gerando um gráfico automático a fim de facilitar a visualização. A utilização de tal recurso permitiu visualizar as áreas de pesquisas em que os resultados se distribuem, onde é possível notar que a Ciência da Computação (83,8%), Matemática (19,3%) e Engenharia (13,5%) são as áreas de mais destaque, respectivamente, conforme a Figura 1.

FIGURA 1 - ÁREAS DE PESQUISA DA BASE SCOPUS- *DATA MINING* E *OPEN DATA*



FONTE: AUTORA COM BASE NOS RESULTADOS DA *SCOPUS* (2017).

WEB OF SCIENCE - Obteve-se 70 resultados combinando as palavras "*Data mining*" e "*Open data*" na base de dados *Web of Science*. Constatou-se que o inglês também é a linguagem predominante, representando 97,14% do resultado da respectiva busca (equivalente a 68 de 70 resultados). Outros idiomas, tal como o alemão e espanhol possuem apenas um cada, que juntos representam 2,84% do total. Assim como na *Scopus*, a *WoS* permite analisar os resultados diretamente em sua página web conforme os parâmetros selecionados, gerando um gráfico automático a fim de facilitar a visualização. Tal recurso também foi utilizado e permitiu visualizar as áreas de pesquisas em que os resultados se distribuem, onde é possível notar que a Ciência da Computação (65,7%), Engenharia (34,3%) e Informática Médica (7,1%) são as áreas de mais destaque, respectivamente, conforme a Figura 2.

FIGURA 2 - ÁREAS DE PESQUISA DA BASE *WEB OF SCIENCE - DATA MINING* E *OPEN DATA*

Campo: Áreas de pesquisa	Contagem do registro	% de 70	Gráfico de barras
COMPUTER SCIENCE	46	65.714 %	
ENGINEERING	24	34.286 %	
MEDICAL INFORMATICS	5	7.143 %	
ENVIRONMENTAL SCIENCES ECOLOGY	3	4.286 %	
TELECOMMUNICATIONS	3	4.286 %	
AUTOMATION CONTROL SYSTEMS	2	2.857 %	
GEOGRAPHY	2	2.857 %	
HEALTH CARE SCIENCES SERVICES	2	2.857 %	
MATHEMATICAL COMPUTATIONAL BIOLOGY	2	2.857 %	
MATHEMATICAL METHODS IN SOCIAL SCIENCES	2	2.857 %	
MATHEMATICS	2	2.857 %	
PHARMACOLOGY PHARMACY	2	2.857 %	
ROBOTICS	2	2.857 %	
ACOUSTICS	1	1.429 %	
BUSINESS ECONOMICS	1	1.429 %	
EDUCATION EDUCATIONAL RESEARCH	1	1.429 %	
ENERGY FUELS	1	1.429 %	
GEOLOGY	1	1.429 %	
GOVERNMENT LAW	1	1.429 %	
INFORMATION SCIENCE LIBRARY SCIENCE	1	1.429 %	
NEUROSCIENCES NEUROLOGY	1	1.429 %	
OCEANOGRAPHY	1	1.429 %	
OPERATIONS RESEARCH MANAGEMENT SCIENCE	1	1.429 %	
PUBLIC ADMINISTRATION	1	1.429 %	
PUBLIC ENVIRONMENTAL OCCUPATIONAL HEALTH	1	1.429 %	
SCIENCE TECHNOLOGY OTHER TOPICS	1	1.429 %	
SOCIOLOGY	1	1.429 %	
Campo: Áreas de pesquisa	Contagem do registro	% de 70	Gráfico de barras

FONTE: AUTORA COM BASE NOS RESULTADOS DA *WEB OF SCIENCE* (2017).

Os resultados apresentados anteriormente corroboram com a afirmação de que há poucas publicações científicas de mineração de dados abertos no idioma em português, com apenas um resultado na base *Scopus* - afirmação que esta fundamenta e justifica o desenvolvimento do presente trabalho.

1.3 OBJETIVOS

A seguir será apresentado o objetivo principal do presente trabalho e os objetivos específicos, que juntos contribuem para o alcance do objetivo geral.

1.3.1 Objetivo Geral

O objetivo geral do presente trabalho constitui-se em aplicar técnicas de mineração, do tipo classificação, nos dados de pessoas com deficiência (PcD) da região Sul do Brasil, provenientes da base pública CAGED (Cadastro Geral de Empregados e Desempregados) dos anos de 2009 à 2016, buscando por padrões que descrevam o mercado de trabalho formal e o perfil das pessoas com deficiente na referida região.

1.3.2 Objetivos Específicos

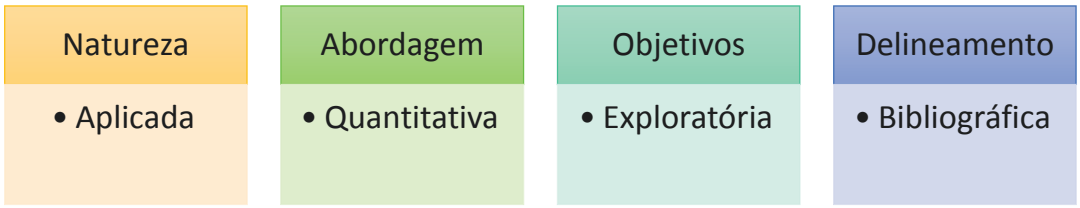
Os objetivos específicos que desdobram o objetivo geral, são:

a) Aplicar técnicas de mineração, do tipo classificação, na base de dados de deficiente da região Sul do Brasil, provenientes do CAGED;
b) Identificar os melhores algoritmos para a referida base de dados, com base nos seus desempenhos;
c) Descrever o mercado de trabalho formal das pessoas com deficiências na região Sul por meio da aplicação de algoritmos de classificação sobre a referida base de dados.
d) Analisar e apresentar os padrões obtidos da mineração, apontando aplicações destes no cotidiano.

1.4 METODOLOGIA DE DESENVOLVIMENTO DA PESQUISA

A presente pesquisa é caracterizada segundo as perspectivas da natureza, abordagem, objetivos e delineamento, apresentados na Figura 3.

FIGURA 3 – CARACTERIZAÇÃO METODOLÓGICA DA PESQUISA



FONTE: A AUTORA (2017).

De acordo com Silva e Menezes (2005, p. 20), há dois tipos de naturezas de pesquisa: pesquisa básica e pesquisa aplicada. A presente pesquisa mais se aproxima da natureza aplicada, que segundo Silva e Menezes (2005, p. 20), visa "gerar conhecimentos para aplicação prática e dirigidos à solução de problemas específicos. Envolve verdades e interesses locais." Gil (1989, p. 44) corrobora com Silva e Menezes ao afirmar que a característica fundamental das pesquisas aplicadas é o "interesse na aplicação, utilização e consequências práticas do conhecimento."

Segundo Silva e Menezes, as pesquisas podem ser classificadas em qualitativa ou quantitativa quanto a sua abordagem. A presente pesquisa é classificada como quantitativa, pois, segundo Silva e Menezes (2005, p. 20), tal classificação "considera que tudo pode ser quantificável, o que significa traduzir em números opiniões e informações para classificá-las e analisá-las."

Segundo Gil (1989) há diferentes níveis de pesquisa a partir do objetivo da mesma: Pesquisas exploratórias, pesquisas descritivas e pesquisas explicativas. Dentre as três apresentadas, a presente pesquisa assume a característica de exploratória, pois, segundo Gil (1989, p. 45), este tipo de pesquisa tem como objetivo proporcionar uma visão aproximada sobre um determinado fato, que no caso específico, refere-se ao mercado formal de trabalho dos deficientes brasileiros.

Já sob o ângulo do seu delineamento, ou seja, dos procedimentos técnicos utilizados para coleta de dados, há diferentes tipos de classificações segundo Gil (1989): pesquisa bibliográfica, pesquisa documental, pesquisa experimental, pesquisa *expost-facto*, levantamentos e estudo de caso. Gil (1989, p. 70) explica que o delineamento da pesquisa “refere-se ao planejamento da pesquisa em sua dimensão mais ampla, envolvendo tanto a sua diagramação quanto a previsão de análise e interpretação dos dados [...]” e complementa em outras palavras que “[...] considera o ambiente em que são coletado os dados, bem como a forma de controle das variáveis envolvidas. Ainda de acordo com Gil (1989, p. 71), tal decisão decorre do procedimento adotado para a coleta de dados, o elemento mais importante para a identificação de um delineamento. Diante disso, a presente pesquisa assume o caráter de pesquisa bibliográfica, que Gil (2008, p. 50) a define como “desenvolvida a partir de material já elaborado, constituído principalmente de livros e artigos científicos [...]” e a presente pesquisa tem como exclusiva a fonte bibliográfica, conforme destacado pelo autor (1989, p. 71; 2008, p. 50).

Utilizou-se o ciclo de gerenciamento da informação de autoria da Beal (2007) como auxilio para as etapas do KDD (*Knowledge Discovery in Databases*), tais como seleção dos dados; limpeza, pré-processamento e transformação dos dados; análise exploratória e seleção de modelos; mineração de dados; interpretação e avaliação; descoberta do conhecimento - Tais etapas foram a base metodológica para o desenvolvimento prático do trabalho, algoritmos dos tipos *tree* e *rules* da tarefa de classificação foram selecionados para aplicação via ferramenta *WEKA* (*Waikato Environment for Knowledge Analysis*).

1.5 ESTRUTURA DE ORGANIZAÇÃO DA PESQUISA

O presente trabalho está organizado em 5 capítulos, que são expostos da seguinte forma:

O Capítulo 1 apresenta a contextualização do problema, a justificativa, os objetivos (geral e específicos) e a metodologia de desenvolvimento.

O Capítulo 2 mostra fundamentação teórica da pesquisa, fazendo uma explanação de temas como gestão da informação; KDD (*Knowledge Discovery in Databases*) e o seu respectivo processo; a mineração de dados, bem como o seu histórico, suas tarefas, técnicas e ferramentas; dados abertos, trazendo aspectos

como a Lei de Acesso à Informação (LAI), o Portal Brasileiro de Dados Abertos (PBDA) e o CAGED (objeto de estudo); as pessoas com deficiência (PcD) e o seu mercado de trabalho no Brasil.

O Capítulo 3 trata da estatística descritiva dos dados a serem minerados.

O Capítulo 4 traz os aspectos práticos da pesquisa, onde detalha-se por etapa do KDD: a seleção, pré-processamento e limpeza dos dados; transformação dos dados; análise exploratória dos dados e seleção de modelos; mineração e interpretação e avaliação.

E finalmente, o Capítulo 5 mostra as considerações finais, com relação aos resultados encontrados, valida os objetivos específicos determinados, e contribuição do presente trabalho, além de também trazer sugestões para pesquisas futuras.

2 REFERENCIAL TEÓRICO

A fim de respaldar teoricamente a presente pesquisa, a seguir são aludidos os seguintes assuntos: gestão da informação (GI), *Knowledge Discovery in Databases* (KDD), mineração de dados, dados abertos, Lei de Acesso à Informação (LAI) e Pessoas com Deficiência (PcD).

2.1 DADO, INFORMAÇÃO E CONHECIMENTO: GESTÃO DA INFORMAÇÃO

A fim de possibilitar e facilitar a compreensão dos assuntos abordados na presente pesquisa, se faz necessário apresentar a distinção entre dado, informação e conhecimento.

No campo da gestão da informação (GI), vários autores dissecam e estabelecem uma relação entre dado, informação e conhecimento, porém, como uma breve introdução ao tema, destacam-se aqui algumas destas definições, inclusive a do Davenport (1998), um dos principais e renomados autores da área, que os distingue conforme apresentado na seguinte Figura 4:

FIGURA 4 - CONCEITO: DADO, INFORMAÇÃO E CONHECIMENTO

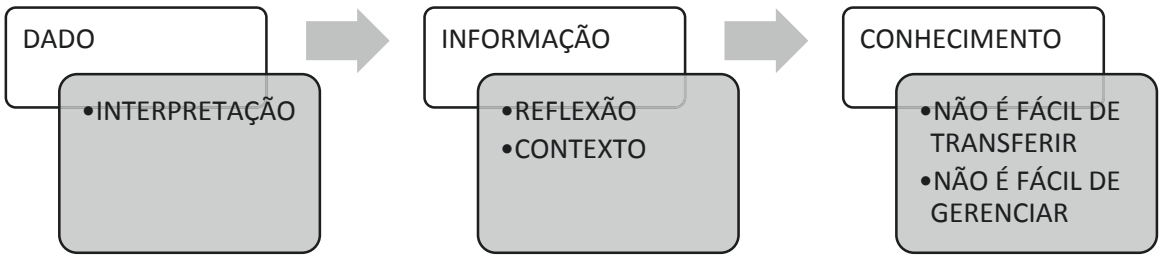
Dados	Informação	Conhecimento
Simples observações sobre o estado do mundo Facilmente estruturado • Facilmente obtido por máquinas • Freqüentemente quantificado • Facilmente transfcrível	Dados dotados de relevância e propósito • Requer unidade de análise • Exige consenso em relação ao significado • Exige necessariamente a mediação humana	Informação valiosa da mente humana Inclui reflexão, síntese, contexto • De difícil estruturação • De difícil captura em máquinas • Freqüentemente tácito • De difícil transferência

FONTE: DAVENPORT (P. 18, 1998).

Para Tarapanoff (2006, p. 23), a informação compreende recursos originados na “produção de dados, tais como de registros e arquivos, que vêm da gestão de pessoal, pesquisa de mercado, da observação e análise utilizando os princípios da inteligência competitiva, de uma vasta gama de fontes.”

Por sua vez, Drucker, outro renomado autor, destaca que “[...] informações são dados dotados de relevância e propósito” (DRUCKER apud DAVENPORT, 1998, p. 19). Conforme destaque na Figura 5 a seguir, é imprescindível uma interpretação dos dados para que se transformem em informações, e, quando a informação compreende reflexão e contexto, possui valor para a mente e não é fácil transferi-la e gerenciar, estamos tratando do conhecimento, segundo Davenport (1998).

FIGURA 5 - DADO, INFORMAÇÃO E CONHECIMENTO



FONTE: ADAPTADO DE DAVENPORT (1998).

No atual contexto, a informação torna-se um dos fatores mais importantes no ambiente competitivo das organizações, sendo considerado um dos principais componentes para manter o nível de competitividade, juntamente com o conhecimento. Além de ser “um fator determinante para a melhoria de processos, produtos e serviços, tendo valor estratégico [...]” (Tarapanoff, p. 23, 2006), a informação também torna-se essencial para qualquer setor da economia, seja ele público ou privado, além de ser intrínseca à outra área de atividade humana e indispensável a qualquer ação, conforme apontam os autores McGee e Prussak (1994):

> [...]nos últimos 25 anos o mundo industrializado vem enfrentando a transição de uma economia industrial para uma economia de informação, e nas próximas décadas, a informação, mais do que a terra ou o capital, será a força motriz na criação de riquezas e prosperidade. (MCGEE; PRUSSAK, 1994, p. 3).

Inomata (2012, p. 41) corrobora ao dizer que "a informação está presente, como insumo, em todas as atividades de uma organização, seja na tomada de decisão, no desenvolvimento de produtos dentre outros processos". Os autores Porém, Santos e Belluzo também enfatizam a importância da informação nos dias atuais:

> No ambiente competitivo e de rápidas mudanças que as organizações enfrentam, a informação é essencial para sua sobrevivência, mas não a informação em si, em sua forma física e estática, e sim a gestão da informação e de seu fluxo a fim de gerar conhecimento e consequentemente oferecer subsídios para as tomadas de decisões nas empresas. (PORÉM, SANTOS E BELLUZO, 2012, p. 01)

Devido a globalização, o avanço das tecnologias e a expansão da inclusão tecnológica, "há um volume cada vez maior de informação, disponibilizada num intervalo de tempo cada vez menor", o que leva à necessidade de gerenciar esse recurso que, “embora abundante [...], tende,” a ser “utilizado de forma ineficiente”. (MARCHIORI, 2002, p. 79; MONTEZANO, 2009, p. 8). Diante disso,

> A Gestão da Informação surge, então, para sanar essa ineficiência, aliando conceitos da Gestão Estratégica às Tecnologias de Informação, nas empresas, com o objetivo de sistematizar e organizar o conhecimento, os dados e as informações (MONTEZANO,2009, p. 9)

Em sua concepção, Davenport (p. 173, 1998) define gestão da informação (GI) como “um conjunto estruturado de atividades que incluem o modo como as empresas obtêm, distribuem e usam a informação e o conhecimento.” Citando Wilson (1997), Tarapanoff (p. 21, 2006) define gestão da informação como “a aplicação de princípios administrativos à aquisição, organização controle, disseminação e uso da informação para a operacionalização efetiva de organizações de todos os tipos”.

Segundo Inomata, na literatura são apontados vários modelos distintos de fluxos ou gerenciamento da informação, mas reforça que cada um "está relacionado ao contexto organizacional, ambiental e informacional, particular de cada segmento." (INOMATA, 2012, p. 52). Um dos modelos bastante difundido na literatura é o do

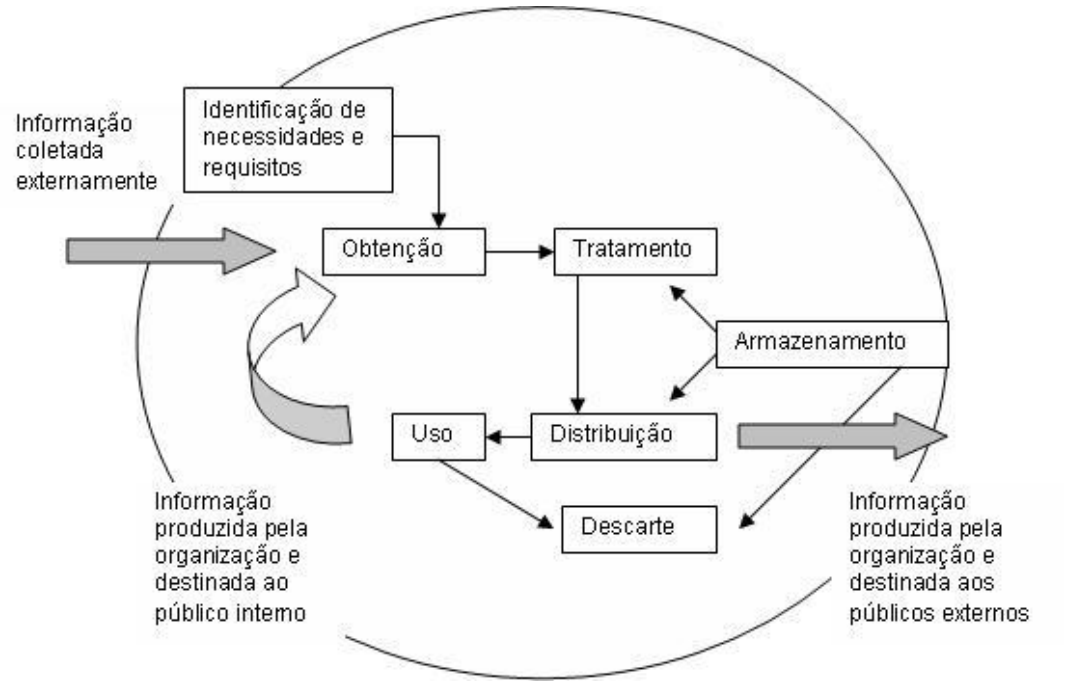
Davenport (p. 175, 1998), apresentado na Figura 6 a seguir, composto pelas fases de 1) determinação das exigências, 2) obtenção, 3) distribuição, 5) utilização:

FIGURA 6 - CICLO DE GESTÃO DA INFORMAÇÃO SEGUNDO DAVENPORT

FONTE: ADAPTADO DE DAVENPORT (p. 175, 1998).

Mas de todos os quais apresenta (Leitão, 1985; Lesca; Almeida, 1994; Navarro, 2000; Forza; Salvador, 2001; Barreto, 2002; Choo, 2003; Beal, 2007), Inomata (2012, p. 64) aponta o modelo da autoria de Beal (2007) como o mais completo, por contemplar o descarte, etapa necessária quando a informação se torna inútil. Tal modelo é usado como referência para fins da presente pesquisa (Figura 7) e detalhado posteriormente:

FIGURA 7 - CICLO DE GESTÃO DA INFORMAÇÃO SEGUNDO BEAL (2007)



FONTE: BEAL (2007, p. 29).

a) **Identificação de necessidades e requisitos:** consiste na identificação das necessidades (do grupo ou do indivíduo) e requisitos de informação, o qual ao serem atendidos são recompensados (BEAL, 2007, p. 30).
b) **Obtenção:** são desenvolvidas as atividades de criação, recepção ou captura de informação, provenientes de fonte interna ou externa em qualquer mídia ou formato (BEAL, 2007, p. 30).
c) **Tratamento:** Beal descreve o tratamento como "processo de organização, formatação, estruturação, classificação, análise, síntese e apresentação, com o propósito de torná-la mais acessível e fácil de localizar pelos usuários" (BEAL, 2007, p. 30).
d) **Distribuição:** prevê o direcionamento da informação necessária a quem precisa, neste processo, "quanto melhor a rede de comunicação da organização, mais eficiente é a distribuição interna da informação", além disso, como ressalta a autora, normalmente a organização precisa preocupar-se com os processos de distribuição de informação para públicos externos, referindo-se a distribuição para o mercado, como parceiros, fornecedores, clientes, acionistas, governo etc. (BEAL, 2007, p. 31).
e) **Uso:** considerada a etapa mais importante pela autora, à medida que o uso da informação possibilita a combinação de informações e o surgimento de novos conhecimentos, que podem voltar a alimentar o ciclo da informação corporativo, num processo contínuo de aprendizado e crescimento. (BEAL, 2007, p. 31).
f) **Armazenamento:** "é necessária para assegurar a conversão dos dados e informações permitindo seu uso e reuso dentro da organização" (BEAL, 2007, p. 31).
g) **Descarte:** o modelo prevê o descarte da informação, obedecendo a normas legais, políticas operacionais e exigências internas ao excluir informações inúteis. Uma vez que melhora o processo de gestão da informação de diversas maneiras, por exemplo, "economizando recursos de armazenamento, aumento a rapidez e eficiência na localização da informação necessária [...]" (BEAL, 2007, p. 31).

O KDD (*Knowledge Discovery in Databases*) é um processo que auxilia a gestão da informação com a extração de informações de base de dados, mostrando

relações de interesse que não são observadas pelo especialista no assunto ou pelo profissional de informação. Conforme citado anteriormente, por outro lado há limitação da capacidade humana de aproveitar o grande volume de dados digitais gerados pelos usos intensivos das tecnologias, onde o KDD mais uma vez auxilia a GI, facilitando a transformação de tais dados em informação útil e conhecimento para suporte à decisão. Suas fases, que serão melhor detalhadas a seguir, complementam as fases de obtenção e tratamento da informação, do modelo proposto por Beal (2007) e por outro lado, a GI em muito auxilia o processo de KDD no que tange a identificação informacional, muito correlacionado ao objetivo de aplicação, também no uso informacional dos padrões e conhecimentos resultados do KDD, além avaliação do desempenho dos resultados para o processo de tomada de decisão.

2.2 KDD - KNOWLEDGE DISCOVERY IN DATABASES

A Descoberta de Conhecimento em Bases de Dados (DCBD) - em tradução literal, surge então da necessidade de ferramentas computacionais, desenvolvidas com o objetivo de ajudar os humanos em sua extração de informações à úteis dos grandes volumes de dados, segundo Fayyad et al (1996).

Independente da área de aplicação, o método clássico de obter informações úteis a partir dos dados consiste na análise e interpretação manual, além de ser altamente subjetiva, tal abordagem está se tornando inviável devido ao imenso volume de dados que só cresce exorbitantemente.

Diante disso, o KDD “[...] é uma tentativa de resolver um problema que a era da informação digital tornou um fato de vida para todos nós: sobrecarga de dados” *(*Fayyad et al, 1996, p. 38, tradução nossa). Fayyad, Piatetsky-Shapiro e Smyth, (1996) apud Fayyad et al (1996, p.41, tradução nossa) definem KDD como “[...] processo não trivial de identificação de padrões válidos, novos, potencialmente úteis e, finalmente, compreensíveis em dados”. Fayyad et al, explicam que neste contexto:

> [...] os dados são um conjunto de fatos [...] e o padrão é uma expressão em algum idioma que descreve um subconjunto dos dados ou um modelo aplicável ao subconjunto. Portanto, [...] a extração de um padrão também designa a adequação de um modelo aos dados; estrutura de pesquisa a partir de dados; ou, em geral, fazer qualquer descrição de alto nível de um conjunto de dados. O termo processo implica que o KDD compreende muitas etapas, que envolvem preparação de dados, busca de padrões, avaliação de conhecimento e refinamento, todos repetidos em múltiplas iterações. Por não

> trivial, queremos dizer que alguma pesquisa ou inferência está envolvida; Isto é, não é uma computação direta de quantidades predefinidas como calcular o valor médio de um conjunto de números. (FAYYAD et al, 1996, p.41, tradução nossa)

Ainda segundo os mesmos autores, o KDD progrediu e continua progredindo a partir da interseção entre diferentes áreas de pesquisa, ou seja, multe e interdisciplinaridade, tais como banco de dados, estatística, inteligência artificial, visualização de dados, aprendizado de máquinas, reconhecimento de padrões e computação de alto desempenho, indicados na Figura 8 a seguir, com o objetivo em comum de “extrair conhecimento de alto nível de dados de baixo nível no contexto de grandes conjuntos de dados” (FAYYAD et al, 1996, p.39, tradução nossa).

FIGURA 8 - INTERDISCIPLINARIDADE DO KDD.

FONTE: FAYYAD ET AL, 1996.

Um exemplo de tal interseção citado pelos autores é a etapa de mineração de dados do KDD, onde diferentes campos - aprendizado de máquina, reconhecimento de padrões e estatísticas - contribuem com métodos usados na referida etapa para encontrar os padrões, porém, segundo os autores, o KDD diferencia destes por se concentrar no processo global de descoberta de conhecimento, desde

> [...] como os dados são armazenados e acessados, como os algoritmos podem ser escalados para conjuntos de dados maciços e ainda funcionar de forma eficiente, como os resultados podem ser interpretados e visualizados e

> como a máquina- Interação pode ser utilmente modelada e suportada. O processo KDD pode ser visto como uma atividade multidisciplinar que engloba técnicas além do escopo de qualquer disciplina particular [...], o KDD coloca uma ênfase especial em encontrar padrões compreensíveis que podem ser interpretados como conhecimento útil ou interessante. (FAYYAD et al, 1996, p.39, tradução nossa)

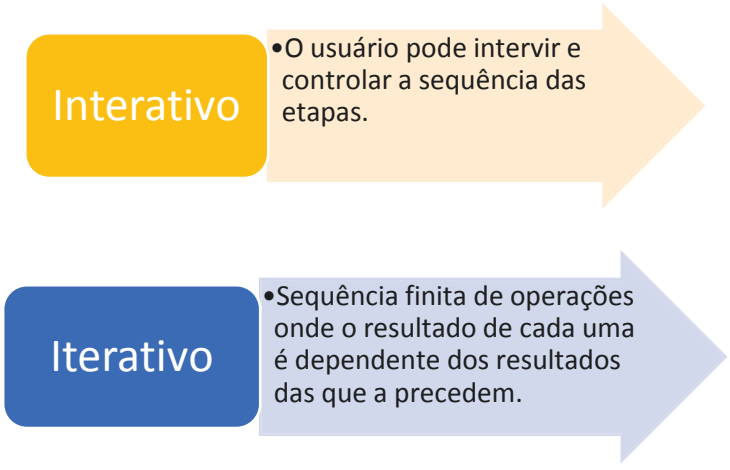
A seguir será detalhado as etapas básicas do KDD, à luz dos autores Brachman e Anand (1996), citados por Fayyad et al (1996).

2.2.1 Processo do KDD

Conforme mencionado anteriormente, o processo interativo e iterativo do KDD envolve o banco de dados juntamente com as etapas de preparação dos dados, seleção dos dados, pré-processamento e limpeza, transformação, mineração de dados e interpretação e avaliação.

Prass (2004, p. 15), explica que o processo assume ambas características porque o usuário pode intervir e controlar a sequência das etapas - interativo, e também porque é uma sequência finita de operações onde o resultado de cada uma é dependente dos resultados das que a precedem - iterativo, conforme indica a Figura 9 a seguir:

FIGURA 9 - CARACTERÍSTICAS DA MINERAÇÃO DE DADOS



FONTE: PRASS (2004).

Com base em outros autores (Adriaans e Zantige, 1996; Brachman e Anand, 1996; Fayyad et al., 1996; Han e Kamber, 2000). Soares

Junior e Quintella (2005) dizem que as etapas do KDD podem ser reunidas em três fases: preparação, análise e interpretação:

a) **Identificação dos objetivos:** embora não presente na Figura 10, em primeiro lugar, autores enfatizam que deve-se identificar o objetivo da aplicação e desenvolver a compreensão do domínio do seu conhecimento. Ainda segundo os mesmos autores, “os objetivos [...] são definidos pelo uso pretendido do sistema”. (FAYYAD et al, 1996, p.43, tradução nossa). Os atores classificam os objetivos em verificação e descoberta e explicam que “[...] com a verificação, o sistema se limita a verificar a hipótese do usuário. Com a descoberta, o sistema autônomo encontra novos padrões”. (FAYYAD et al, 1996, p.43, tradução nossa).
b) **Seleção:** etapa que consiste em “selecionar um conjunto de dados ou focar um subconjunto de variáveis ou amostras de dados, em que a descoberta deve ser realizada”. (FAYYAD et al, 1996, p.42, tradução nossa).
c) **Limpeza de dados e pré-processamento:** terceira etapa, que conforme o nome diz, consiste na limpeza dos dados. Como operações básicas

> Incluem remoção de ruído, se apropriado, coletando as informações necessárias para modelar ou contabilizar o ruído, decidir sobre estratégias para manipular campos de dados ausentes e contabilizar informações de sequência de tempo e mudanças conhecidas.” (FAYYAD et al, 1996, p.42, tradução nossa).

d) **Transformação:** etapa que consiste na “redução e projeção de dados: encontrar recursos úteis para representar os dados dependendo do objetivo da tarefa. Com métodos de redução de dimensionalidade ou de transformação, o número efetivo de variáveis sob consideração pode ser reduzido, ou podem ser encontradas representações invariantes para os dados.” (FAYYAD et al, 1996, p.42, tradução nossa)
e) Embora não presente na Figura 10, a presente etapa consiste em combinar os objetivos do processo KDD (etapa a) com um método particular de mineração de dados, cujos métodos podem ser classificação, regressão, agrupamento e etc. (FAYYAD et al, 1996, p.42, tradução nossa)
f) **Análise exploratória e seleção de modelos e hipóteses:** embora não presente na Figura 10, nesta etapa é a escolha da tarefa de mineração de

dados e do (s) algoritmo (s) a serem usados para pesquisar padrões de dados. Segundo Fayyad et al

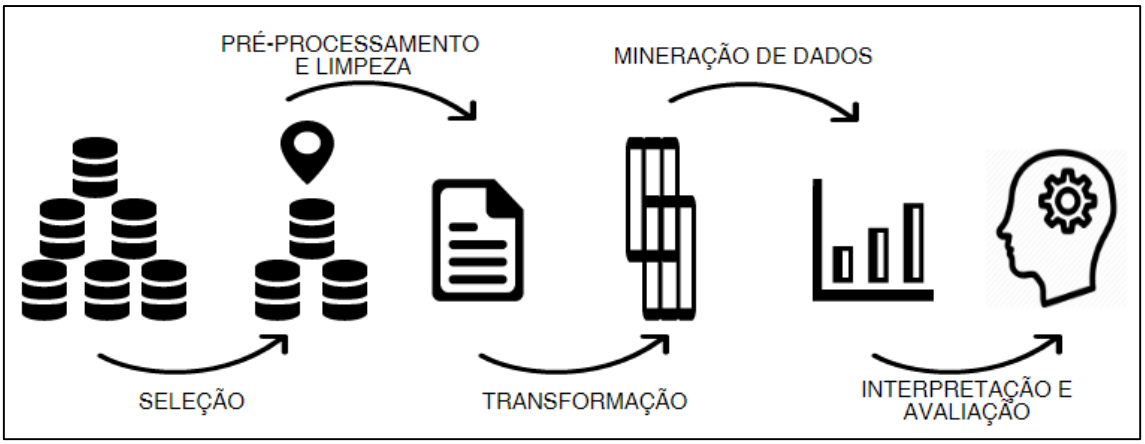
> Este processo inclui decidir quais modelos e parâmetros podem ser apropriados (por exemplo, os modelos de dados categóricos são diferentes dos modelos de vetores sobre os reais) e combinando um método particular de mineração de dados com os critérios gerais do processo KDD. (FAYYAD et al, 1996, p.42, tradução nossa)

g) **Mineração de dados:** etapa onde ocorre a mineração de dados, propriamente dita, pois, é a "busca de padrões de interesse em uma forma representacional particular ou um conjunto de tais representações, incluindo regras de classificação ou árvores, regressão e agrupamento." (FAYYAD et al, 1996, p.42, tradução nossa).
h) **Interpretação e avaliação:** etapa onde consiste em interpretar

> Padrões minados, possivelmente retornando a qualquer um dos passos 1 a 7 para iteração adicional. Esta etapa também pode envolver a visualização dos padrões e modelos extraídos ou visualização dos dados. (FAYYAD et al, 1996, p.42, tradução nossa).

i) **Descoberta do conhecimento:** consiste em agir sobre o conhecimento descoberto: usando o conhecimento diretamente, incorporando o conhecimento em outro sistema para ação futura, ou simplesmente documentando-o e informando-o às partes interessadas. Este processo também inclui a verificação e resolução de potenciais conflitos com conhecimento previamente acreditado (ou extraído). (FAYYAD et al, 1996, p.42, tradução nossa).

FIGURA 10 - ETAPAS DO PROCESSO DO *KNOWLEDGE DISCOVERY IN DATABASES (KDD).*



FONTE: ADAPTAÇÃO DE FAYYAD *ET AL.* (1996, P. 41).

Muitos trabalhos se concentram apenas na etapa de mineração de dados, porém, Fayyad et al ressaltam que “as outras etapas são tão importantes ou talvez mais, para a aplicação bem-sucedida do KDD na prática” (1996, p. 42, tradução nossa), além de serem cruciais para garantir que o conhecimento útil seja derivado dos dados. Na visão dos autores Fayyad et al (1996, p.39), a mineração de dados é apenas uma etapa dentro do processo KDD, que por sua vez, refere-se ao processo geral de descoberta de conhecimento útil a partir de dados.

2.3 MINERAÇÃO DE DADOS

Do inglês “*Data Mining*”, mineração de dados é a exploração de quantidade significativa de dados para identificar padrões e regras. De acordo com os autores Fayyad et al (1996, p.39, tradução nossa), “mineração de dados é a aplicação de algoritmos específicos para extrair padrões a partir de dados” e “[...] envolve a adequação de modelos ou a determinação de padrões a partir de dados observados” *(*1996, p.43, tradução nossa). Os mesmos autores complementam que

> A mineração de dados é um passo no processo KDD que consiste em aplicar algoritmos de análise e descoberta de dados que, sob limitações de eficiência computacional aceitáveis, produzem uma enumeração particular de padrões (ou modelos) sobre os dados. (FAYYAD et al, 1996, p.41, tradução nossa).

Berry e Linoff corroboram com Fayyad et al quando definem mineração de dados como “mineração de dados, como usamos o termo, é a exploração e análise

de grandes quantidades de dados, a fim de descobrir padrões e regras significativas." (2004, p.7, tradução nossa). Fayyad et al, (1996, p.43, tradução nossa) acrescentam que

> Os modelos montados desempenham o papel do conhecimento inferido: se os modelos refletem conhecimento útil ou interessante é parte do processo KDD global e interativo onde o julgamento humano subjetivo é normalmente requerido" (FAYYAD et al,1996, p.43, tradução nossa)

Conforme os mesmos autores, a lógica e a estatística são dois formalismos matemáticos primários usados no encaixe do modelo e explicam que "a abordagem estatística permite efeitos não-determinísticos no modelo, enquanto que um modelo lógico é puramente determinista." (FAYYAD et al, 1996, p.43, tradução nossa).

2.3.1 Histórico de mineração de dados

De acordo com Soares Junior e Quintella (2005) o KDD ou DCBD - Descoberta de Conhecimento em Bases de Dados surgiu no final da década de 1980, como um novo ramo da computação, devido a necessidade de extrair sistematicamente o conhecimento e de se produzir novos conhecimentos das bases de dados, cujo volume crescia rapidamente. Segundo os autores (2005, p. 1083), "o objetivo principal de encontrar uma maneira estruturada de, com o uso da TI, explorar essas bases de dados e reconhecer os padrões existentes pela modelagem de fenômenos do mundo real".

Segundo Fayyad et al (1996, tradução nossa), o entendimento de encontrar padrões úteis em grande volume de dados recebeu uma variedade de nomes historicamente, incluindo desde mineração de dados, extração de conhecimento, descoberta de informações, coleta de informações, arqueologia de dados até processamento de padrões de dados. Porém, o termo que ganhou popularização na área de banco de dados e tem sido amplamente usado por estatísticos, analistas de dados e as comunidades de sistemas de informação de gestão (SIG) é o de mineração de dados. Já nas áreas de estudo da Inteligência Artificial (AI) e aprendizagem de máquina o termo mais popular é a frase "Descoberta de Conhecimento em Bancos de Dados", em sua tradução literal, que foi cunhada no primeiro *workshop* de KDD em 1989 (Piatetsky-Shapiro, 1991 apud Fayyad et al, 1996, p. 39, tradução nossa) com o objetivo de enfatizar que o conhecimento é o produto final de uma descoberta baseada em dados.

De acordo com Soares Junior e Quintella (2005, p. 1084), o primeiro registro descritivo dos processos de KDD data de 1996 e é de autoria dos pesquisadores Usama Fayyad, Gregory Piatetsky Shapiro e Padhraic Smyth do Massachusetts Institute of Technology (MIT), cujo artigo é intitulado "*The KDD process for extracting useful knowledge from volumes of data*". No referido artigo, os autores demonstram sua preocupação em sistematizar as etapas do processo KDD, já que segundo os próprios

> A maioria dos trabalhos anteriores sobre o tema dava ênfase à etapa de *Data mining*. No entanto, os outros passos são igualmente, se não mais, importantes para o sucesso da aplicação de KDD na prática (FAYYAD et al, 1996, p. 39)

Especificamente sobre mineração de dados, no prefácio do seu livro "Introdução à Mineração de Dados" (2016), Castro e Ferrari nos informam que

> Surgiu como área de pesquisa e aplicação independente em meados da década de 1990, mas suas origens na matemática, estatística e computação são muito anteriores a esse período. A área também ganhou evidência nos últimos anos depois de ser cunhado o termo *Big Data* e com a publicação do relatório intitulado *Big Data: The Next Frontier for Innovation, Competition, and Productivity* pelo McKinsey Global Institute em meados de 2011. (CASTRO E FERRARI, 2016, p. 13)

Tendo o seu histórico sucintamente apresentado, na subseção a seguir apresenta-se as principais tarefas de mineração de dados.

2.3.2 Tarefas de mineração de dados

De acordo com Castro e Ferrari (2016, p. 41), as funcionalidades da mineração de dados são usadas para especificar os tipos de informações a serem obtidas nas tarefas de mineração e que no geral, tais tarefas podem ser classificadas em duas categorias: descritivas e preditivas, sendo que a primeira caracterizam as propriedades gerais dos dados; e a segunda fazem inferência a partir dos dados, objetivando predições. Segundo Fayyad et al (1996, tradução nossa) a previsão e a descrição visam ser na prática, os dois objetivos primários de alto nível da mineração de dados e corroboram com Castro e Ferrari ao dizerem que

> [...] a previsão envolve o uso de algumas variáveis ou campos no banco de dados para prever valores desconhecidos ou futuros de outras variáveis de interesse, e a descrição se concentra em encontrar padrões interpretáveis por humanos descrevendo os dados. (FAYYAD et al, 1996, p.43, tradução nossa)

Segundo os autores, a importância de cada objetivo pode variar consideravelmente, pois, são relativa e de acordo com aplicações específicas, mas que, seja qual for a meta, ambas podem ser alcançadas usando uma variedade de técnicas de mineração de dados específicos ou divisão dos algoritmos de mineração de dados.

Especialmente a técnica classificação, cujos detalhes apresenta-se a seguir, “é a tarefa preditiva de identificação da classe à qual um objeto pertence” (CASTRO e FERRARI, 2016, p. 258).

Segundo Carvalho (2001, p. 21), alguns autores apresentam um número muito grande de técnicas básicas, porém, “apenas cinco abraçam didaticamente todas as outras formas de apresentação”, permitindo uma visão global e apropriada para uma introdução ao assunto: classificação, estimativa, previsão, análise de afinidade e análise de agrupamentos, apresentadas respectivamente a seguir:

Classificação (*classification*) - De uma forma simplificada, a classificação é uma função que mapeia ou classifica um item de dados em uma das várias classes predefinidas (WEISS e KULIKOWSKI 1991; HAND 1981 apud FAYYAD et al, 1996, p.44, tradução nossa). Berry e Linoff (2004, p. 9), complementam que a classificação “consiste em examinar as características de um objeto recém-apresentado e atribuí-lo a um conjunto de classes predefinido.” Eis alguns exemplos práticos citados pelos autores (BERRY e LINOFF, 2004; FAYYAD et al, 1996):

a) A classificação de tendências nos mercados financeiros;
b) A identificação automatizada de objetos de interesse em bases de dados de grandes imagens;
c) Classificar candidatos de crédito como risco baixo, médio ou alto;
d) Seleção de conteúdo a ser exibido em uma página da Web;
e) Determinar quais números de telefone correspondem a máquinas de fax;
f) Detecção de reclamações de seguro fraudulentas;
g) Atribuição de códigos de indústria e designações de cargos com base em descrições de trabalho em texto livre.

Estimação (*estimation*) ou regressão (*regression*) - Segundo Fayyad et al (1996, p.44, tradução nossa), a regressão consiste numa função que mapeia um item de dados para uma variável de predição de valor real. Berry e Linoff (2004, p.9, tradução nossa) complementam que diferente da classificação, cujos resultados são discretos: sim ou não; feminino e masculino, etc., a estimação lida com resultados continuamente valorizados.

Dado alguns dados de entrada, estimativa vem acima com um valor para alguma variável contínua desconhecida tal como a renda, a altura, ou o contrapeso do cartão de crédito, onde a tarefa de classificação se resume ao estabelecimento de uma pontuação limiar, por exemplo: Qualquer pessoa com uma pontuação maior ou igual ao limiar é classificada como alta, e qualquer pessoa com uma pontuação mais baixa é considerada baixa. (BERRY e LINOFF, 2004, p. 9, tradução nossa). Os autores ainda trazem exemplos de aplicações práticas (BERRY e LINOFF, 2004; FAYYAD et al, 1996):

a) Estimativa do número de crianças em uma família;
b) Estimativa da renda familiar total de uma família;
c) Estimativa do valor de vida de um cliente;
d) Estimativa da probabilidade de que alguém responda a uma solicitação de transferência de saldo;
e) Predizer a quantidade de biomassa presente em uma floresta dada medições de microondas de sensoriamento remoto;
f) Estimar a probabilidade de que um paciente sobreviva dado os resultados de um conjunto de testes diagnósticos.

Previsão ou predição (*prediction*) - Diferente da classificação ou estimativa, na tarefa de predição os registros são classificados de acordo com algum comportamento futuro previsto ou valor futuro estimado. A principal razão para tratar a predição como uma tarefa distinta da classificação e estimativa é que na modelagem preditiva existem questões adicionais sobre a relação temporal das variáveis de entrada ou preditores com a variável alvo, segundo Berry e Linoff (2004, p. 10, tradução nossa). Segundo os autores Berry e Linoff (2004, p. 10, tradução nossa), qualquer uma das técnicas utilizadas para classificação e estimativa pode ser adaptada para uso na previsão, usando exemplos de treinamento onde o valor da variável a ser predita já é

conhecido, juntamente com dados históricos para esses exemplos. Os autores explicam dados históricos como

> Os dados históricos são usados para construir um modelo que explique o comportamento observado atual. Quando este modelo é aplicado às entradas atuais, o resultado é uma previsão de comportamento futuro. (BERRY E LINOFF, 2004, p. 10, tradução nossa)

Os exemplos práticos de aplicação que os autores citam são:

a) Previsão do tamanho do saldo que será transferido se um cliente de cartão de crédito aceita uma oferta de transferência de saldo
b) Previsão de quais clientes sairão nos próximos 6 meses
c) Prever quais assinantes de telefone solicitarão um serviço de valor agregado, como chamadas de três vias ou correio de voz

Análise de afinidade ou associação de regras (*association*) - Também conhecida por agrupamento de afinidade, sua tarefa é determinar quais coisas vão juntas, ou seja, “O agrupamento de afinidade é uma abordagem simples para gerar regras a partir de dados.” (BERRY e LINOFF, 2004, p. 10, tradução nossa). O exemplo clássico é a análise da cesta de mercado, onde pode determinar o que as coisas vão juntos em um carrinho de compras no supermercado. As redes de varejo podem usar o agrupamento de afinidade para planejar o arranjo de itens nas prateleiras das lojas ou em um catálogo para que os itens comprados juntos sejam vistos juntos. Outro exemplo de uso citado pelos autores seria a identificação de oportunidades de venda cruzada e para projetar pacotes atrativos ou agrupamentos de produtos e serviços. Se dois itens, digamos comida de gato e maca de gatinho, ocorrem juntos com frequência suficiente, podemos gerar duas regras de associação (BERRY e LINOFF, 2004, p. 10, tradução nossa):

a) As pessoas que compram alimentos para gatos também compram lixo com probabilidade P1.
b) Pessoas que compram maca de gatinho também comprar comida para gatos com probabilidade P2.

Análise de agrupamentos (*clustering*) - O agrupamento implica em identificar um conjunto finito de categorias ou clusters para descrever os dados (JAIN e DUBES

1988; TITTERINGTON, SMITH e MAKOV, 1985 apud FAYYAD et al, 1996, p.45, tradução nossa). Berry e Linoff complementam que “agrupamento é a tarefa de segmentar uma população heterogênea em um número de subgrupos mais homogêneos ou clusters.” (2004, p. 11, tradução nossa) e ainda enfatizam que a diferença com a classificação é que o grupo ou cluster não depende de classes pré-definidas:

> Na classificação, a cada registro é atribuída uma classe predefinida com base em um modelo desenvolvido através de treinamento em exemplos pré-classificados. No agrupamento, não existem classes predefinidas e não existem exemplos. Os registros são agrupados com base na auto-similaridade. Cabe ao usuário determinar qual o significado, se houver, para anexar aos clusters resultantes. (BERRY E LINOFF, 2004, p. 11, tradução nossa)

Exemplos de aplicações, segundo os autores (BERRY e LINOFF, 2004; FAYYAD et al, 1996), incluem:

a) A descoberta de subpopulações homogêneas para consumidores em bases de dados de marketing;
b) A identificação de subcategorias de espectros de medições de céu infravermelho;
c) Clusters de sintomas podem indicar doenças diferentes;
   Clusters de atributos de clientes podem indicar diferentes segmentos de mercado.

2.3.3 Técnicas de mineração de dados

Especialmente sobre a tarefa de classificação de dados, apresentada anteriormente, atualmente uma variedade de técnicas ou heurísticas, que servem para diferentes propósitos “está fortemente relacionada com o tipo de conhecimento que se deseja extrair ou com o tipo de dado no qual ela será aplicada” (PRASS, 2004, p. 20). As quais Castro e Ferrari (2016) nos apresentam, são:

BASEADOS EM CONHECIMENTO - Segundo os autores, esse tipo de classificador opera por meio de um conjunto de regras usadas para atribuir determinada classe a um objeto caso ele satisfaça condições predefinidas. Os autores apresentam a seguinte Figura 11 como representação gráfica do modelo:

FIGURA 11 - REPRESENTAÇÃO GRÁFICA DOS ALGORITMOS BASEDOS EM CONHECIMENTO

FONTE: CASTRO E FERRARI (2016, P. 269)

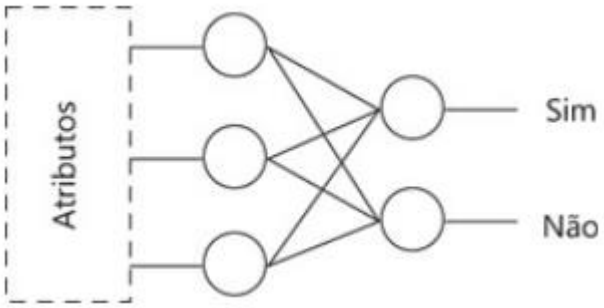
ÁRVORES DE DECISÃO - Árvore de decisão é um modelo preditivo, cujo resultado pode ser visualizado na forma de uma árvore: o nó raiz e os nós intermediários das árvores representam testes sobre um atributo, os ramos representam os resultados desses testes e os nós folhas, os rótulos de classe. Os autores apresentam a seguinte Figura 12 como representação gráfica do modelo:

FIGURA 12- REPRESENTAÇÃO GRÁFICA DOS ALGORITMOS DE ÁRVORE DE DECISÃO

FONTE: CASTRO E FERRARI (2016, P. 269)

REDES NEURAIS ARTIFICIAIS - Também conhecidos como conexionistas, são aqueles modelos baseados em redes de unidades (nós) interconectadas. Os sistemas conexionistas são um tipo de grafo e, embora haja diferentes sistemas conexionistas, os mais comuns são as Redes Neurais Artificiais (CASTRO e FERRARI, 2016, p. 270). Estruturalmente, a referida técnica compõe-se de um número de elementos interconectados, os chamados neurônios, organizados em camadas que aprendem pela modificação de suas conexões. Tipicamente, tem-se uma camada de entrada ligada a uma ou mais camadas intermediárias que são ligadas a uma camada de saída (BERRY e LINOFF, 1997 apud PRASS, 2004, p. 22). A função básica de cada neurônio é “avaliar valores de entrada, calcular o total para valores de entrada combinados, comparar o total com um valor limiar e determinar o valor de saída.” (PRASS, 2004, p. 22). Castro e Ferrari apresentam a seguinte Figura 13 como representação gráfica do modelo:

FIGURA 13 - REPRESENTAÇÃO GRÁFICA DOS ALGORITMOS DE REDES NEURAIS



FONTE: CASTRO E FERRARI (2016, P. 270)

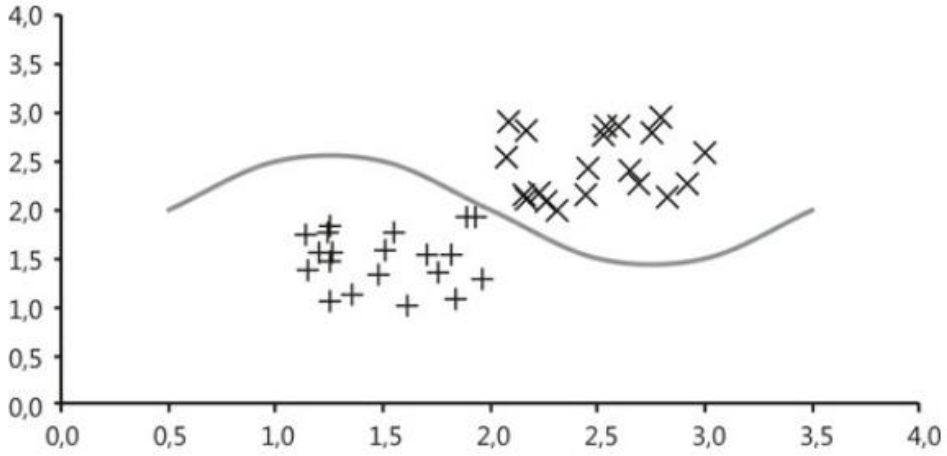
BASEADOS EM DISTÂNCIA: Segundo os autores Castro e Ferrari, em tais modelos o processo de classificação se dá calculando a distância entre o objeto cuja classe se deseja conhecer e um ou mais objetos rotulados. A classe do objeto desconhecido passa a ser a mesma daqueles objetos que estão a uma menor distância dele. Os autores apresentam a seguinte Figura 14 como representação gráfica do modelo:

FIGURA 14 - REPRESENTAÇÃO GRÁFICA DOS ALGORITMOS BASEADOS NA DISTÂNCIA

FONTE: CASTRO E FERRARI (2016, P. 270)

BASEADOS EM FUNÇÃO: São modelos paramétricos baseados em funções predefinidas e cujos parâmetros são ajustados durante o processo de treinamento. Após o treinamento, um novo objeto de classe desconhecida é apresentado à função, cujo valor é calculado e que representa, de alguma forma, a classe desse objeto. Os autores apresentam a seguinte Figura 15, como representação gráfica do modelo:

FIGURA 15 - REPRESENTAÇÃO GRÁFICA DOS ALGORITMOS BASEADOS EM FUNÇÃO



FONTE: CASTRO E FERRARI (2016, P. 271)

PROBABILÍSTICOS: Permitem atribuir uma probabilidade de um objeto pertencer a uma ou mais classes possíveis. Os autores apresentam a seguinte Figura 16 como representação gráfica do modelo:

FIGURA 16- REPRESENTAÇÃO GRÁFICA DOS ALGORITMOS PROBABILÍSTICOS

FONTE: CASTRO E FERRARI (2016, P. 271)

Porém, Castro e Ferrari ainda enfatizam que tais categorias de algoritmos não são mutuamente exclusivas, pois, “um modelo baseado em função pode ser probabilístico, assim como um classificador do tipo árvore também pode ser interpretado como um sistema baseado em regras.” (CASTRO e FERRARI, 2016, p. 272).

2.3.4 Ferramentas de mineração de dados

Atualmente há uma variedade de ferramentas ou *software*s para aplicação de mineração de dados, tanto privadas quanto de códigos abertos - *software*s livres, que “foram desenvolvidas no intuito de tornar a aplicação da Mineração de Dados uma tarefa menos técnica, e com isto possibilitar que profissionais de outras áreas possam fazer usa dela” (CAMILO e SILVA, 2009, p. 21).

Em seu livro intitulado “Introdução à Mineração de Dados” (2016), da editora Saraiva, Castro e Ferrari trazem um capítulo especial só para tais ferramentas, destacando os mais utilizados, dos quais algumas são indicadas a seguir:

a) ELKI - O ELKI (*Environment for Developing KDD-Applications Supported by Index-Structures*) é um *software* de mineração de dados de código aberto (AGPLv3) desenvolvido em Java e mantido pela Universidade de Munique Ludwig-Maximilians (elki.dbs.ifi.lmu.de), de acordo com os autores.
b) LIBSVM - O LIBSVM é uma biblioteca gratuita para desenvolvimento de Máquinas de Vetores de Suporte (*Support Vector Machines* - SVM) desenvolvida em Java e C++, e mantida por uma comunidade de pesquisadores (www.csie.ntu.edu.tw/~cjlin/libsvm). De acordo com os autores, as SVMs são técnicas que podem ser aplicadas para tarefas de otimização, classificação, regressão, estimação, entre outras.
c) MAHOUT - De acordo com os autores, é um *software* gratuito de código aberto (*Apache License*) mantido por uma comunidade de desenvolvedores. O Mahout é um projeto da Apache e já conta com algoritmos para agrupamento, classificação e recomendação que podem ser aplicados em base de dados de larga escala (*Big Data*) com processamento paralelo ou distribuído.
d) MATLAB - Matlab é uma linguagem de programação de alto nível aliada a um ambiente de desenvolvimento pago (www.mathworks.com/products/matlab). Segundo os autores, Atualmente a ferramenta conta com pacotes para diferentes processos de mineração de dados, tais como agrupamento, classificação e estimação, com ferramentas específicas para dados financeiros e biológicos, processamento de sinais e imagens, entre outros.
e) ORANGE - O Orange é um *software* gratuito baseado na linguagem de programação Python. Permite a construção visual, por meio de blocos e

fluxogramas, de processos complexos de análise e mineração de dados, segundo os autores. Possui também pacotes adicionais para atuar nas áreas de bioinformática, mineração de textos e visualização de dados.

f) R - O R é uma linguagem e ambiente gratuito para computação estatística e visualização de dados (gráficos), de código aberto (*GNU General Public License*) mantido por uma comunidade de pesquisadores e desenvolvedores (www.rproject.org). Segundo os autores, em virtude de sua origem estatística, foi adotado por diferentes grupos de pesquisas que acabaram por criar pacotes de algoritmos para as mais diversas tarefas de mineração de dados.

g) RAPIDMINER - O RapidMiner é um *software* que atua em processos de mineração de dados e possui versões gratuitas e pagas (rapidminer.com). O *software* permite a construção visual, por meio de blocos e fluxogramas, de processos complexos de análise e mineração de dados, podendo conectar-se a diferentes fontes de dados, tais como arquivos e diferentes SGBDs, de acordo com os autores.

h) SAS - Segundo os autores, a empresa SAS (*Statistical Analysis System*) possui uma gama de *software*s para gerenciamento de bases de dados, análise preditiva, mineração de dados e visualização de dados (www.sas.com), entre eles o *Enterprise Miner*, que segundo Amorin (2006, p. 37), o *software* “integra diferentes técnicas da mineração de dados, sendo uma avançada ferramenta para predição e descrição de dados, utilizando diversos algoritmos”, tais como os de árvores de decisão, redes neurais e outros.

i) *WEKA* - Uma das ferramentas que vem recebendo destaque - principalmente no meio acadêmico, o *WEKA (Waikato Environment for Knowledge Analysis),* é um *software* de código aberto emitido sob a *GNU General Public License*. Desenvolvida pela *The University of Waikato*, o *WEKA* é uma coleção de algoritmos de aprendizagem de máquina para tarefas de mineração de dados, que contém ferramentas para pré-processamento de dados, classificação, regressão, agrupamento, regras de associação e visualização (http://www.cs.waikato.ac.nz/ml/*WEKA*/index.html).

Para aplicação do presente trabalho, optou-se pela ferramenta WEKA, que além das características citadas anteriormente, a escolha também deve-se pelo fato da mesma possuir as seguintes características: disponibilidade de grande quantidade de algoritmos da tarefa de classificação, visualização dos resultados graficamente

(dependendo do algoritmo, pode ser gráfico ou árvores), possibilidade de configuração dos parâmetros dos algoritmos e diferentes opções de acesso aos dados, principalmente a possibilidade de ser direto do Banco de Dados (SQL Server, por exemplo).

2.3.5 Tipos de dados

Dados são valores quantitativos ou qualitativos associados a alguns atributos, de uma forma simplificada segundo Castro e Ferrari (2016). Ainda segundo os autores, podem ser classificados quanto a sua estrutura, quanto aos seus valores, quanto a sua dimensionalidade e quanto ao seu formato, conforme mostra a Figura 17 a seguir:

Quanto a sua estrutura, os dados podem ser classificados em:

a) Estruturado: quando residem em campos fixos em um arquivo, por exemplo: planilha ou banco de dados;
b) Semi-estruturado: não possui a estrutura completa de um modelo de dados, mas também não é totalmente desestruturado, por exemplo: XML;
c) Não-estruturado: são aqueles não organizados de uma maneira predefinida ou que não residem em locais definidos, por exemplo: sons, páginas da web, imagens, arquivos PDFs, etc.

Já o valor, que é a medida da quantidade do dado, podem ser classificados em numérico ou categórico.

a) Numérico: podem assumir quaisquer valores numéricos, que por sua vez podem ser classificados em discretos ou contínuos:
    - Discretos: valores inteiros
    - Contínuos: valores reais
b) Categórico: quando assumem valores correspondentes a símbolos distintos, que por sua vez podem ser binários, nominais ou ordinais, introduzidos como níveis de medidas pela estatística:
    - Binário: aqueles que assumem apenas dois valores possíveis, por exemplo “0” ou “1”

- Nominal: aquele cujos valores possuem símbolos ou rótulos distinto, por exemplo: o atributo “estado civil” pode assumir os valores solteiro”, “casado”, “separado”, “divorciado” e “viúvo”
- Ordinal: aquele que permite ordenar suas categorias, embora não necessariamente haja uma noção explícita de distância entre as categorias, por exemplo: pré; fundamental, médio, superior e pós
- Razão: quantidades do tipo razão são aquelas para as quais o método de medida define o ponto zero. Por exemplos, peso, distância, velocidade, salário etc.

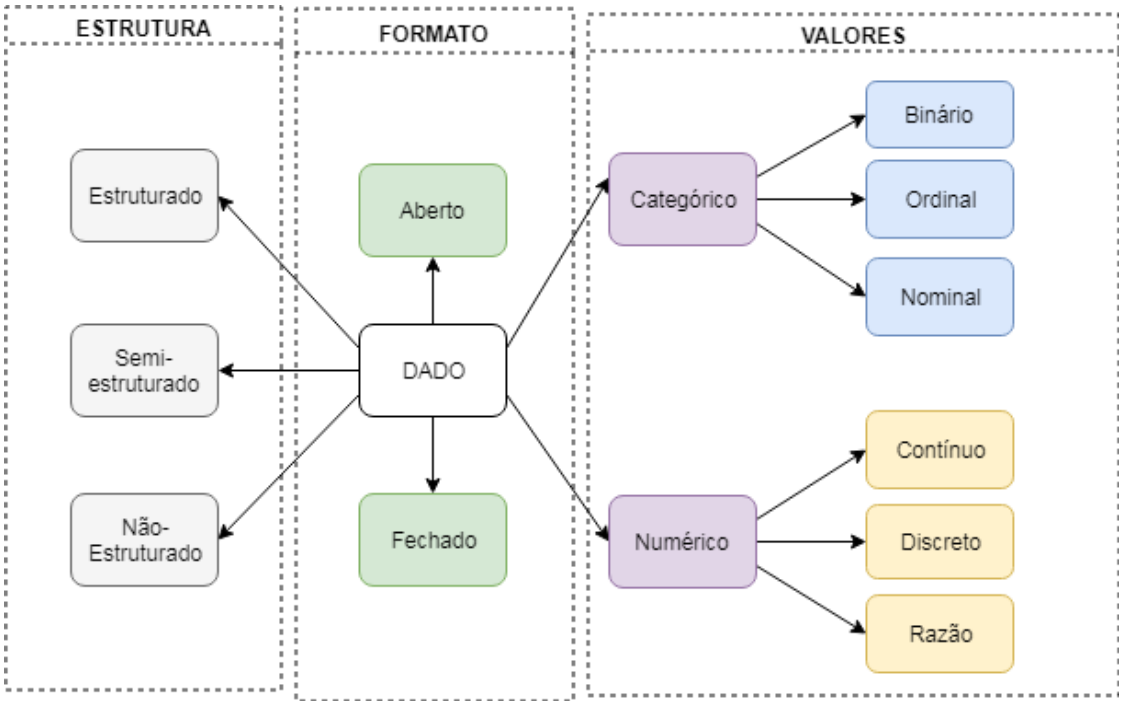
Quanto a sua dimensionalidade, os dados podem ser classificados em uni variadas, bivariadas e multivariadas:

a) Uni variadas:
b) Bivariadas:
c) Multivariadas:

Quanto ao formato de publicação ou armazenamento, os dados podem ser fechados ou abertos. O formato fechado é quando as especificações não estão disponíveis publicamente, ou que, mesmo com as especificações abertas ao público, a reutilização é limitada (MANUAL DOS DADOS ABERTOS: GOVERNO, 2011).

Com mais detalhes a seguir na seção a seguir, ainda segundo o mesmo manual, o formato aberto é quando as especificações do *software* são abertas para qualquer pessoa ver, é gratuito e que pode ser aberto por diferentes tipos de programas, sem limitação de reutilização imposta por direitos de propriedade intelectual.

FIGURA 17 - TIPO DE DADOS



FONTE: A AUTORA (2017).

2.4 DADOS ABERTOS

Apesar de não estar claro para a maioria das pessoas, *Big Data* e *Open Data* são assuntos intrinsecamente relacionados.

Do inglês, *Open Data*, Dados Abertos é a publicação e disseminação dos dados e informações públicas na Internet, no formato aberto e sob licença aberta, organizados de tal maneira para que toda e qualquer pessoa pode acessá-los, manuseá-los e compartilhá-los, independente da finalidade, desde que preservem sua abertura e proveniência. Máchová e Lněnička definem dados abertos como,

> [...] um pedaço de conteúdo ou dados que alguém é livre para usar, reutilizar e também redistribuí-los - sujeitos apenas, no máximo, à exigência de atribuir e compartilhar. A maioria dos dados abertos está na verdade em forma bruta. No entanto, a republicação implica citar a fonte original não apenas para dar crédito, mas para garantir que esses dados não tenham sido modificados ou deturpados. (MÁCHOVÁ E LNĚNIČKA, 2017, p 24)

Valdivia, Navarrete e Aracena, por sua vez, complementam que dados abertos são

> [...] dados provenientes de ministérios, municípios, províncias, municípios, universidades e escolas, entre outros, estão associados a uma grande variedade de informações (bens de capital "público" compartilhado) que podem ser de interesse para muitos usuários. (VALDIVIA, NAVARRETE E ARACENA, 2014, p. 2, tradução nossa)

Para Estermann (2014, p. 16, tradução nossa), o termo inclui todos os tipos de dados, tais como "relatórios de estudos, mapas, fotos de satélite, fotos e pinturas, dados meteorológicos, dados geográficos e ambientais, dados da pesquisa, o genoma, dados médicos, ou fórmulas científicas."

Ainda segundo Estermann, o movimento de dados abertos teve origem nos círculos acadêmicos há mais de 50 anos, e teve um avanço no mundo todo quando

> A administração Obama e o Governo do Reino Unido adotaram políticas de Dados Governamentais Abertos, a fim de promover a transparência, participação e colaboração entre políticos, autoridades públicas, empresas privadas e cidadãos. (ESTERMANN, 2014, p. 16, tradução nossa)

Garcia (2014, p. 82, tradução nossa), cita as várias razões do porque os dados abertos são importantes:

a) Fornecer um controle externo à corrupção [...] e mau uso do dinheiro público, bem como a avaliação da economia, eficiência e eficácia.
b) Eles fazem as organizações abrir seus dados mais eficiente e eficaz através de entidades de participação cidadã e de colaboração.
c) Setor privado pode desenvolver novos usos e aplicações para a informação, [...] com valor econômico.
d) Melhorar a tomada de decisão dos indivíduos com base em informações que não estavam disponíveis anteriormente.

Já para Valdivia, Navarrete e Aracena (2014, p. 458, tradução nossa), a "importância dos dados aberto encontra-se em um conjunto de razões que descrevem o *Open Knowledge Foundation"*:

1. Transparência: Para que uma sociedade democrática funcione corretamente, os cidadãos precisam saber o que seu governo está fazendo. Eles, [a sociedade] devem ser capazes de acessar livremente dados do governo para compartilhar esta informação e cruzar com informações de outros cidadãos.
2. Fornecer valor social e comercial: Na era digital, os dados são um recurso fundamental para atividades social e comercial. Através de dados abertos, os

governos podem ajudar [...] na criação de empresas e serviços inovadores que oferecem valor social e comercial.

3. Governo Participativo: Através de dados aberto, os cidadãos têm direito a ser mais diretamente informados e envolvidos na tomada de decisões. Isso é mais do que transparência, é sobre a construção de uma sociedade justa, não só por ser ciente do que acontece no processo de governo, mas ser capaz de contribuir para isso.

Diante disso, se faz necessário uma introdução à Lei de Acesso à Informação, visto que está intrinsecamente relacionada com os dados abertos e deste é impossível desvincular, pois, em art. 8º determina que

> É dever dos órgãos e entidades públicas promover, independentemente de requerimentos, a divulgação em local de fácil acesso, no âmbito de suas competências, de informações de interesse coletivo ou geral por eles produzidas ou custodiadas." (LAI, 2017).

2.4.1 Lei de acesso à informação - LAI

O acesso a informação está previsto na Declaração Universal dos Direitos Humanos, definida em 1948 e na Constituição Federal de 1988, porém, Jardim (2012) informa que

> Foram necessários 23 anos para que o Brasil contasse com uma Lei de Acesso à Informação Pública que favorecesse a aplicação dos princípios do direito à informação presentes na Constituição de 19881, apesar do tema ter sido contemplado - embora jamais implementado - no artigo 5º da chamada Lei de Arquivos de 1991. (JARDIM, 2012, p. 2).

Segundo Jardim (2012, p. 2), "90 países já haviam adotado legislação semelhante" à Lei de Acesso à Informação - LAI, como é popularmente conhecida no Brasil. Sancionada no dia 18 de novembro de 2011, a Lei nº 12.527 instaura o direito constitucional de acesso às informações públicas a todos os cidadãos brasileiros, ou seja, toda e qualquer informação produzida ou sob guarda do poder público. Por meio dela, o governo brasileiro criou "mecanismos que possibilitam, a qualquer pessoa, física ou jurídica, sem necessidade de apresentar motivo, o recebimento de informações públicas dos órgãos e entidades" (LAI, 2017). Apesar de ter sido publicada no dia 18 de Novembro de 2011, a referida norma entrou em vigor somente

no dia 16 de maio de 2012, quando regulamentada pelo Poder Executivo Federal, ou seja, 180 dias após. Para Ventura, a referida legislação

> [...] regulamenta o amplo direito ao acesso à informação pública, determinando deveres estatais de gerir de forma eficiente a documentação governamental ou sob sua guarda, e viabilizar o conhecimento e a consulta a todos. Disponibilidade, autenticidade, integridade são os principais atributos legais da informação pública. (VENTURA, 2013, p. 636).

Há alguns casos específicos que não se enquadram na definição acima, tais como as informações pessoais, as informações sigilosas com base em outras leis (sigilos bancário, fiscal e industrial, por exemplo) e as informações classificadas como sigilosas pelos órgãos competentes - cuja divulgação possa colocar em risco a segurança da sociedade (vida, segurança, saúde da população) ou do Estado (soberania nacional, relações internacionais, atividades de inteligência).

O Decreto 7.724, em seu art. 13, que regulamenta a LAI no Poder Executivo, prevê que não serão atendidos aos cidadão solicitante pedidos genéricos, desproporcionais ou desarrazoados e que exijam trabalhos adicionais - de análise, interpretação ou consolidação de dados e informações, ou serviço de produção ou tratamento de dados que não seja de competência do órgão ou entidade.

A abrangência da lei se entende:

> A Lei vale para os três Poderes da União, Estados, Distrito Federal e Municípios, inclusive aos Tribunais de Conta e Ministério Público. Entidades privadas sem fins lucrativos também são obrigadas a dar publicidade a informações referentes ao recebimento e à destinação dos recursos públicos por elas recebidos. (BRASIL, LEI DE ACESSO À INFORMAÇÃO)

Além da LAI, o governo federal tem trabalhado efetivamente a fim de garantir efetivar a todos os direito à informação, tal como a disponibilização do portal brasileiro de dados abertos, do qual será informado mais detalhes na seção a seguir, dada sua relação direta com uns dos assuntos aqui já tratados.

2.4.2 Portal brasileiro de dados abertos - PBDA

Disponibilizada pelo governo brasileiro com o objetivo de propiciar todo e qualquer dado, o Portal Brasileiro de Dados Abertos (PBDA) é uma ferramenta que disponibiliza aos cidadãos informações e dados públicos, para que possam localizar

e utilizar facilmente, funcionando como um grande catálogo que permeia todos os poderes das esferas federal, estadual e municipal.

> Ele é um serviço simplificado que organiza e padroniza o acesso aos dados públicos, primando pelo reuso dos dados e o uso de tecnologias modernas. [...] O Portal Brasileiro de Dados Abertos organiza os dados abertos em um catálogo para fácil localização pelo cidadão. Entretanto, os responsáveis pelos dados são as organizações públicas que os publicam. Por isso cada uma dessas organizações que responde pelos seus próprios dados. (BRASIL, PORTAL BRASILEIRO DE DADOS ABERTOS).

Atualmente, o portal disponibiliza apenas com uma parcela dos dados publicados pelo governo, porém, segundo consta na sua página web:

> O plano estratégico prevê que nos próximos 3 anos o portal disponibilize acesso aos dados publicados por todos os órgãos do governo federal, além de dados das esferas estaduais e municipais. (BRASIL, PORTAL BRASILEIRO DE DADOS ABERTOS).

Diferente dos portais de transparências, cujo objetivo é o de aumentar o controle das despesas e receitas no governo - da União, dos Estados, do Distrito Federal e dos Municípios, no PBDA é possível acessar uma variedade de dados, tais como: dados da saúde suplementar, do sistema de transporte, de segurança pública, indicadores de educação, gastos governamentais, processo eleitoral, etc.

Além do objetivo de ser uma referência única para a busca e o acesso à dados públicos brasileiros de todo e qualquer assunto ou categoria, o portal também tem o objetivo de “promover a interlocução entre atores da sociedade e com o governo para pensar a melhor utilização dos dados em prol de uma sociedade melhor.”

O portal é um dos compromissos do Brasil perante a *Open Government Partnership* (OGP) ou Parceria para Governo Aberto, no seu primeiro Plano de ação (referenciado pelo Decreto sem número de 15 de setembro de 2011), quando o país foi um dos co-fundadores.

A *Open Government Partnership* (OGP) foi fundada oficialmente no dia 20 de setembro de 2011, sendo uma iniciativa internacional que visa transmitir e incentivar globalmente práticas governamentais relacionadas à transparência dos governos, ao acesso à informação pública e à participação social.

O projeto do portal é coordenado pela Secretaria de Tecnologia da Informação - STI do Ministério do Planejamento, Orçamento e Gestão - MPO e faz parte da Infraestrutura Nacional de Dados Abertos (INDA) e assim como os seus demais

produtos, o PBDA foi desenvolvido com ampla participação da sociedade, contando com a participação de todos os setores da sociedade, incluindo academia, setor privado, órgãos públicos e grupos da sociedade organizada.

2.4.3 CAGED - Cadastro Geral de Empregados e Desempregados

Primitivamente, o CAGED foi “criado como registro permanente de admissões e dispensa de empregados, sob o regime da Consolidação das Leis do Trabalho (CLT)” (Ministério do Trabalho e Previdência Social, 2017).

Instituído pela Lei 4.923 de 23/12/1965 como um instrumento de acompanhamento e fiscalização do processo de admissão e dispensa de trabalhadores regidos pela CLT, atualmente o CAGED é uma das importantes fontes de informação do mercado de trabalho de âmbito nacional e de periodicidade mensal. A partir de 1986, passou a ser utilizado como suporte ao pagamento do seguro-desemprego e, mais recentemente, tornou-se, também, um relevante instrumento à reciclagem profissional e à recolocação do trabalhador no mercado de trabalho (MTPS, 2017).

De periodicidade anual e abrangendo todo o território nacional, o CAGED possui cerca de 905.550 estabelecimentos declarantes mensalmente de acordo com o Ministério do Trabalho (2017). O conjunto de suas informações possibilita o cálculo do índice de emprego, taxa de rotatividade e a flutuação de emprego, desagregados em nível geográfico, setorial e ocupacional. Permite igualmente a obtenção de dados sobre os atributos dos empregados admitidos e desligados: gênero, grau de escolaridade, faixa etária, salários e tempo de emprego.

Atualmente o CAGED se constitui como uma das principais fontes sobre o mercado formal de trabalho no Brasil, servindo como “base para a elaboração de estudos, pesquisas, projetos e programas ligados ao mercado de trabalho, ao mesmo tempo em que subsidia a tomada de decisões para ações governamentais, como por exemplo, é “utilizado pelo Programa de Seguro-Desemprego, para conferir os dados referentes aos vínculos trabalhistas, além de outros programas sociais” (MTPS, 2017).

Seus principais indicadores ou variáveis, são:

a) Flutuação do emprego;
b) Índice de emprego;
c) Taxa de rotatividade;

d) Informações dos estabelecimentos
    - Total dos estabelecimentos informantes
    - Total de admissões
    - Total de desligamentos
    - Taxa de rotatividade
    - Saldo ou variação absoluta do emprego
    - Variação relativa do emprego
    - Admissões por tipo de movimentação (1º emprego, reemprego, reintegração, contrato prazo determinado, transferência)
    - Desligamentos por tipo de movimentação (dispensado, espontâneo, aposentado, morto, transferido)
e) Informações dos empregados
    - Total de admitidos ou desligados por faixa etária
    - Total de admitidos ou desligados por sexo, segundo a faixa etária
    - Total de admitidos ou desligados por faixa etária, segundo o grau de instrução
    - Total de admitidos ou desligados, segundo a ocupação e a remuneração
    - Total de desligados, segundo tempo no emprego e remuneração, etc.

Originalmente, a base de dados do CAGED é composta por 40 atributos ou variáveis, que podem ser consultados na Tabela disponível no APÊNDICE A.

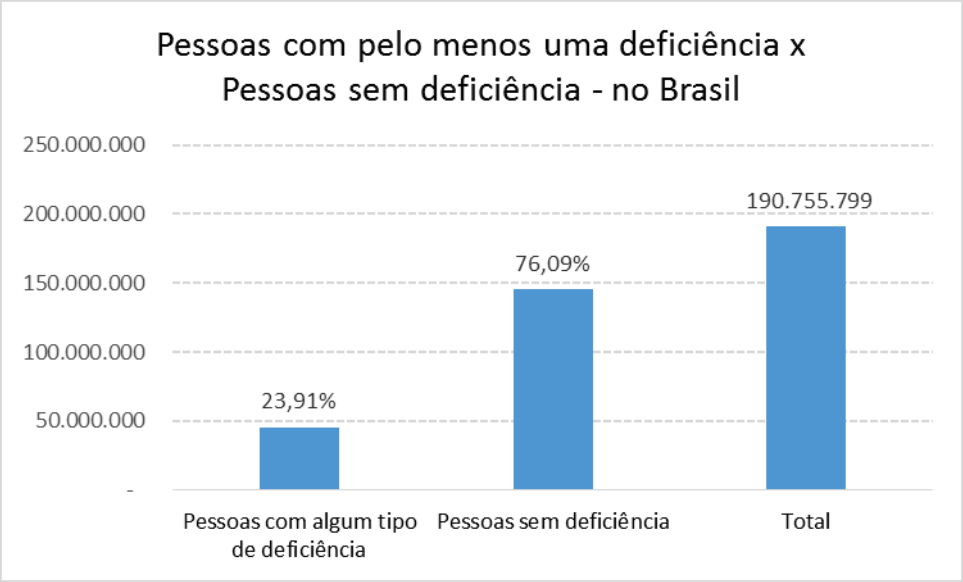
Para fins de mineração, selecionou-se somente os cadastros de pessoas com deficiência da região sul, o qual será dado mais detalhes na seção a seguir. As demais limitações e formas de seleções dos dados serão detalhadas na seção de aplicação, fase seleção de dados - etapa do KDD, conforme apresentado anteriormente.

2.5. PESSOAS COM DEFICIÊNCIA (PcD)

Segundo os dados coletados pelo Instituto Brasileiro de Geografia e Estatística - IBGE, no censo demográfico de 2010, 45.606.048 brasileiros possuíam algum tipo de deficiência - seja deficiência visual, auditiva, motora e mental ou

intelectual, o que representava 23,91% da população total de 190.755.799 habitantes (vide Gráfico 1).

GRÁFICO 1 - POPULAÇÃO RESIDENTE NO BRASIL COM PELO MENOS UMA DEFICIÊNCIA.



FONTE: IBGE (2010).

No Brasil, o Decreto Nº 3.298, de 20 de dezembro de 1999 dispõe sobre a Política Nacional para a Integração da Pessoa Portadora de Deficiência e segundo consta no mesmo, deficiência é:

> Toda perda ou anormalidade de uma estrutura ou função psicológica, fisiológica ou anatômica que gere incapacidade para o desempenho de atividade, dentro do padrão considerado normal para o ser humano (BRASIL. Decreto nº 3.298, 1999, art. 3)

Diferente de pessoa portadora de deficiência habilitada, que segundo o 2º do art. 36 do Decreto nº 3.298/99

> É aquela que concluiu curso de educação profissional de nível básico, técnico ou tecnológico, ou curso superior, com certificação ou diplomação expedida por instituição pública ou privada, legalmente credenciada pelo Ministério da Educação ou órgão equivalente, ou aquela com certificado de conclusão de processo de habilitação ou reabilitação profissional fornecido pelo INSS (BRASIL. Decreto 3.298, 1999, art. 36)

Segundo o Decreto 18 298/1999, somente se o cidadão se encaixar nas categorias indicadas na Figura 18 a seguir - detalhadas posteriormente, é considerado deficiente:

FIGURA 18 - TIPOS DE DEFICIÊNCIA SEGUNDO O DECRETO 3.298/1999

FONTE: A AUTORA (2017).

a) Deficiência física - alteração completa ou parcial de um ou mais segmentos do corpo humano, acarretando o comprometimento da função física, apresentando-se sob a forma de paraplegia, paraparesia, monoplegia, monoparesia, tetraplegia, tetraparesia, triplegia, triparesia, hemiplegia, hemiparesia, ostomia, amputação ou ausência de membro, paralisia cerebral, nanismo, membros com deformidade congênita ou adquirida, exceto as deformidades estéticas e as que não produzam dificuldades para o desempenho de funções;
b) Deficiência auditiva - perda bilateral, parcial ou total, de quarenta e um decibéis (dB) ou mais, aferida por audiograma nas frequências de 500HZ, 1.000HZ, 2.000Hz e 3.000Hz;
c) Deficiência visual - cegueira, na qual a acuidade visual é igual ou menor que 0,05 no melhor olho, com a melhor correção óptica; a baixa visão, que significa acuidade visual entre 0,3 e 0,05 no melhor olho, com a melhor correção óptica; os casos nos quais a somatória da medida do campo visual em ambos os olhos for igual ou menor que 60º; ou a ocorrência simultânea de quaisquer das condições anteriores;

d) Deficiência mental - funcionamento intelectual significativamente inferior à média, com manifestação antes dos dezoito anos e limitações associadas a duas ou mais áreas de habilidades adaptativas, tais como:

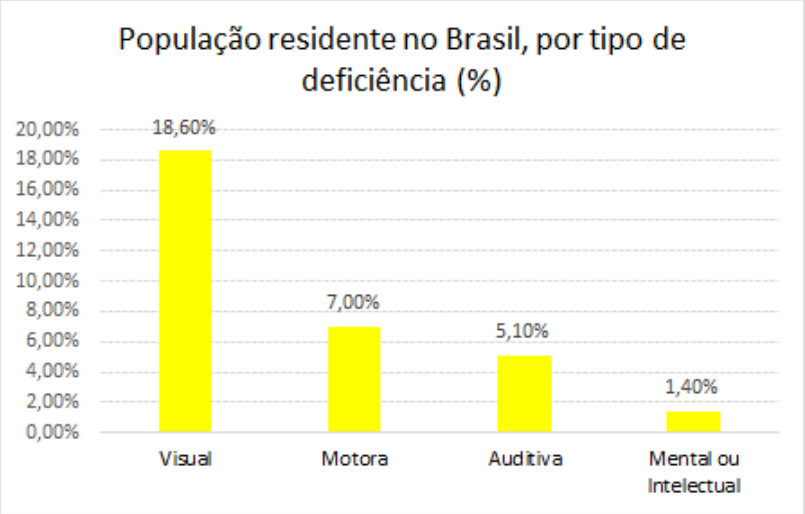
- Comunicação;
- Cuidado pessoal;
- Habilidades sociais;
- Utilização dos recursos da comunidade;
- Saúde e segurança;
- Habilidades acadêmicas;
- Lazer; e
- Trabalho;

e) Deficiência múltipla - associação de duas ou mais deficiências.

De acordo com os dados coletados pelo IBGE em 2010, as categorias de deficiência no Brasil se representam das formas apresentadas a seguir:

Considerando a população residente no país, os deficientes visuais representam 18,60%, os auditivos representam 5,10%, 7% a deficiência motora e por fim, a deficiência mental ou intelectual atingem 1,40% do total de brasileiros (vide Gráfico 2).

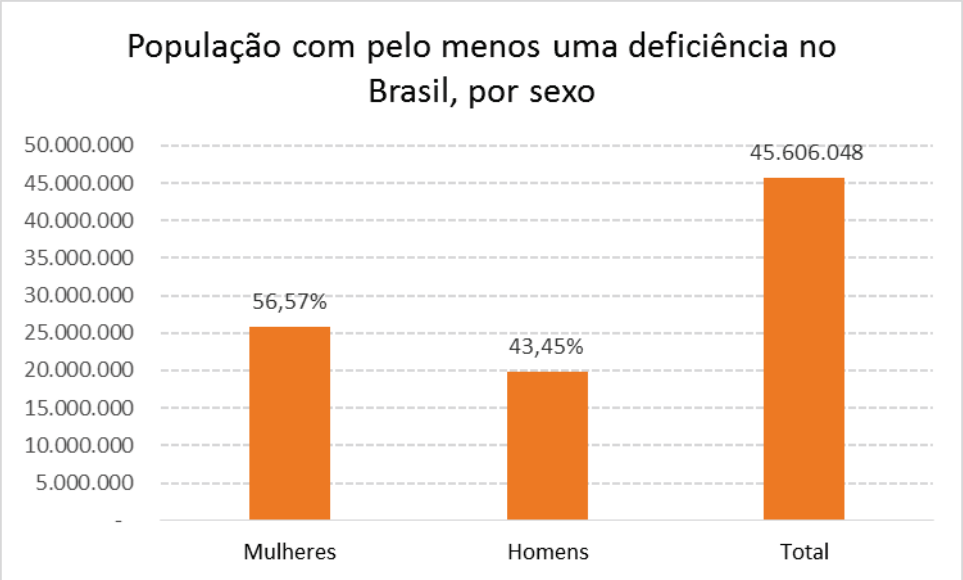
GRÁFICO 2 - POPULAÇÃO RESIDENTE NO BRASIL COM PELO MENOS UMA DEFICIÊNCIA POR TIPO DE DEFICIÊNCIA



FONTE: IBGE (2010).

Do total deficientes (45.606.0480) no Brasil, 56,57% são mulheres (25.800.681) e 43,45% são homens (19.805.367), vide Gráfico 3.

GRÁFICO 3 - POPULAÇÃO RESIDENTE NO BRASIL COM PELO MENOS UMA DEFICIÊNCIA POR SEXO

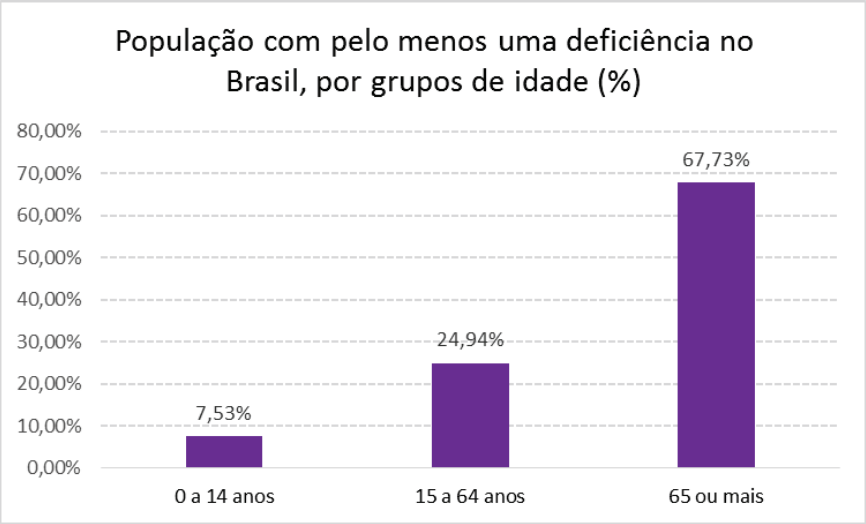


FONTE: IBGE (2010).

Com relação aos grupos de idade no Gráfico 4, o IBGE divulga que com relação ao total de deficientes:

- O grupo de 0 a 14 anos, a deficiência atinge 7,53%;
- 24,9% do grupo de 15 a 64 anos possuem pelo menos uma das deficiências pesquisadas;
- E no grupo de 65 anos ou mais, a mesma relação 67,73%.

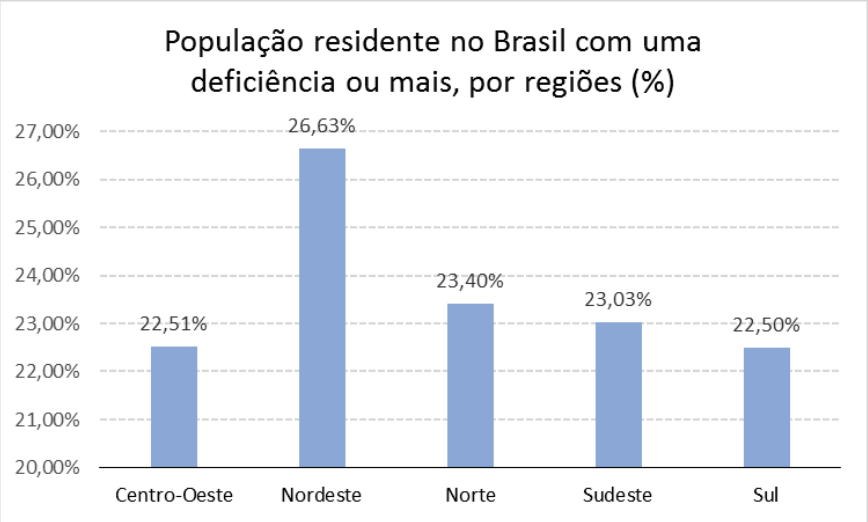
GRÁFICO 4 - POPULAÇÃO RESIDENTE NO BRASIL COM PELO MENOS UMA DEFICIÊNCIA POR GRUPOS DE IDADE



FONTE: IBGE (2010).

Com relação as regiões no Brasil, ou seja, a população geral, a Região Nordeste teve a maior taxa de prevalência de pessoas com pelo menos uma das deficiências, de 26,3%, seguido da Região Norte e Sudeste (com 23,40% e 23,03%, respectivamente). As menores incidências ocorreram nas regiões Sul e Centro Oeste, 22,5% e 22,51%, respectivamente. (Ver Gráfico 5)

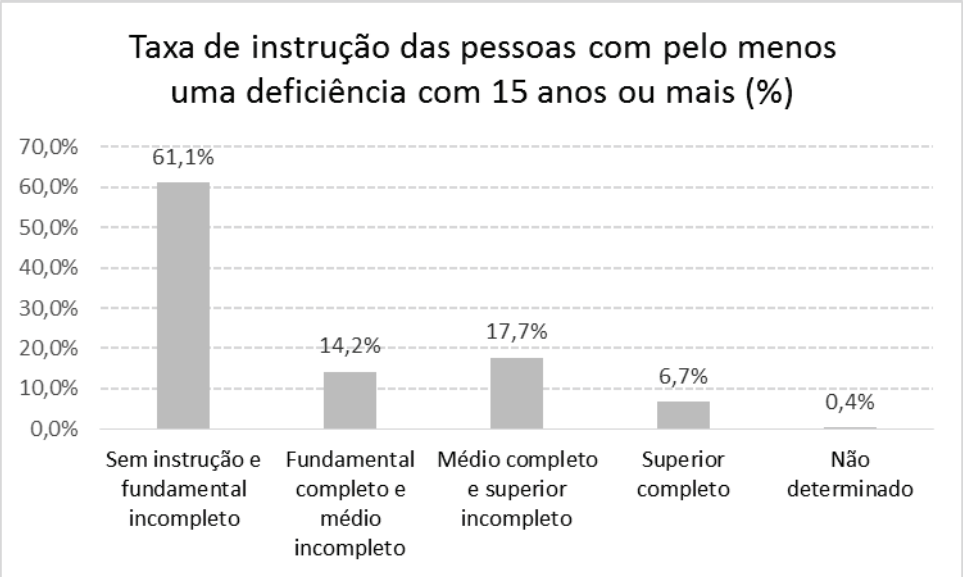
GRÁFICO 5 - POPULAÇÃO RESIDENTE NO BRASIL COM PELO MENOS UMA DEFICIÊNCIA POR REGIÃO



FONTE: IBGE (2010)

Segundo o IBGE, o nível de instrução mede a proporção de pessoas de 15 anos ou mais de idade que atingiram determinados anos de estudo. Para a população com pelo menos uma deficiência, o IBGE (2010) aponta que um total de 61,1% sem instrução e fundamental completo. 14,2% possuíam o fundamental completo, 17,7%, o médio completo e 6,7% possuíam superior completo. A proporção denominada “não determinada” foi igual a 0,4%, conforme consta no Gráfico 6.

GRÁFICO 6 - TAXA DE INSTRUÇÃO DA POPULAÇÃO RESIDENTE NO BRASIL COM PELO MENOS UMA DEFICIÊNCIA COM 15 ANOS OU MAIS

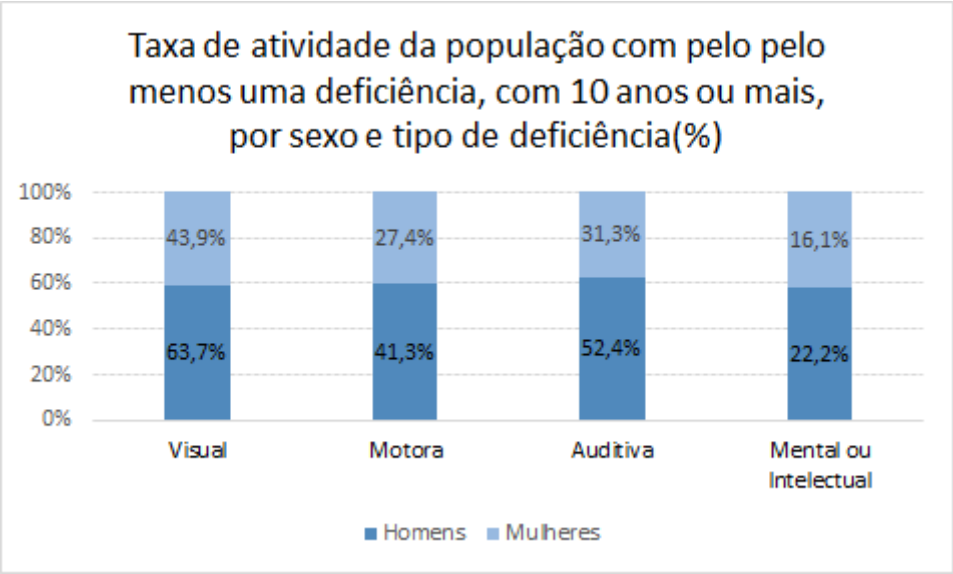


FONTE: IBGE (2010)

Um indicador usado pelo IBGE (2010) para aferir a inserção das pessoas no mercado de trabalho é a taxa de atividade, que mede o percentual de pessoas economicamente ativas na população de 10 ou mais anos de idade. Segundo os dados de 2010, para as pessoas com pelo menos uma das deficiências, essa taxa era de 60,3% para os homens e 41,7% para as mulheres. Quanto aos tipos de deficiência, o grupo das pessoas com deficiência mental ou intelectual recebem destaque por apresentarem a menor taxa, tanto para os homens quanto para as mulheres, cujos valores foram 22,2% e 16,1%, respectivamente e também a menor diferença entre os grupos: 6,1%. Em segundo lugar fica a deficiência motora (com taxas de 41,3% para os homens e 27,4% para mulheres, tendo uma diferença de 13,9%). A deficiência auditiva está em terceiro lugar com 52,4% para homens e 31,3% para mulheres, mas recebe destaque por mostrar a maior diferença entre os grupos: 21,1%. A deficiência

visual foi a menos restritiva, apresentou taxa de 63,7% para homens e 43,9% para mulheres, porém, com uma diferença de 19,8% (Vide Gráfico 7).

GRÁFICO 7 - TAXA DE ATIVIDADE DA POPULAÇÃO RESIDENTE NO BRASIL COM PELO MENOS UMA DEFICIÊNCIA POR SEXO E TIPO DE DEFICIÊNCIA



FONTE: IBGE (2010).

Os direitos humanos são garantidos a todos os cidadãos brasileiros com algum tipo de deficiência e o Governo Federal e da Secretaria Nacional de Promoção dos Direitos da Pessoa com Deficiência promovem programas e ações para esse grupo, porém, as pessoas que apresentam a respectiva deficiência no grau severo são o foco primário de tais políticas públicas.

O conjunto de pessoas identificadas por possuir deficiência severa foi calculado pela soma das respostas positivas às perguntas do IBGE com “tem grande dificuldade” e “não consegue de modo algum”, ou seja, pessoas que responderam que enfrentam “alguma dificuldade” em ouvir, enxergar e em se locomover.

Embora haja uma forte relação entre os dados por faixas etárias de pessoas com pelo menos uma das deficiências investigadas e as pessoas com deficiências severas, a proporção desse último grupo é bem menor do daquelas com pelo menos uma das deficiências: Segundo os dados recolhidos pelo IBGE, em 2010, 8,3% da população brasileira apresentava pelo menos um tipo de deficiência severa, sendo: 0 a 14 anos, a deficiência atinge 7,53% para o primeiro segmento e 2,39% para o

segundo; no grupo de 15 a 64 anos, a relação é de 24,9% e 7,13% e no grupo de 65 anos ou mais, 67,73% e 41,81%.

2.5.1 Mercado de trabalho para pessoas com deficiência

A Declaração Universal dos Direitos Humanos, adotada e proclamada pela Resolução nº 217 A (III) da Assembleia Geral das Nações Unidas em 10 de dezembro de 1948 e assinada pelo Brasil na mesma data, declara, em seu Artigo 23, que “toda pessoa tem direito ao trabalho, à livre escolha do seu trabalho e a condições equitativas e satisfatórias de trabalho e à proteção contra o desemprego”.

Em seu art. 7, a Constituição Federal do Brasil, proíbe a discriminação na remuneração e nos critérios de admissão dos trabalhadores com deficiência e garante a reserva de vagas na administração pública para pessoas com deficiência em seu art. 37.

A Convenção sobre os Direitos da Pessoa com Deficiência, promulgado pelo governo brasileiro via Decreto 6.949/2009, em seu art. 27 trata do trabalho e emprego, onde reafirma o Artigo 23 da Declaração Universal dos Direitos Humanos e assegura, também, condições de acessibilidade que garantam às pessoas com deficiência as mesmas condições de que goza a população sem deficiência.

Com respaldo na Lei nº 7.853/89, em seu art. 8, inciso III, é crime negar a alguém, sem justa causa, emprego ou trabalho em razão de sua deficiência. Além de estar sob pena do inciso XXXI do art. 7º da Constituição Federal, onde proíbe discriminação quanto aos critérios de admissão do trabalhador portador de deficiência.

Outra iniciativa do governo brasileiro que fornece garantias ao trabalhador com deficiência é o Decreto 3.298/1999, popularmente conhecida como a Lei de Cotas, onde no seu art. 36 estabelece que a empresa com 100 ou mais funcionários está obrigada a preencher de dois a cinco por cento dos seus cargos com pessoas com deficiência e reabilitadas, na seguinte proporção do número total de funcionários: até 200, 2%; de 201 a 500, 3%; de 501 a 1.000, 4%; de 1001 e acima, 5%.

É visível que a participação dos trabalhadores no mercado de trabalho é baixa comparada à das pessoas sem deficiência, apesar da exigência legal de cotas para trabalhadores com deficiência. Segundo os dados coletados pelo IBGE em 2010, do total de 86,4 milhões de pessoas, de 10 anos ou mais, ocupadas, apenas 23,6% eram pessoas com deficiência (equivalente a 20,4 milhões do total). Conforme dito

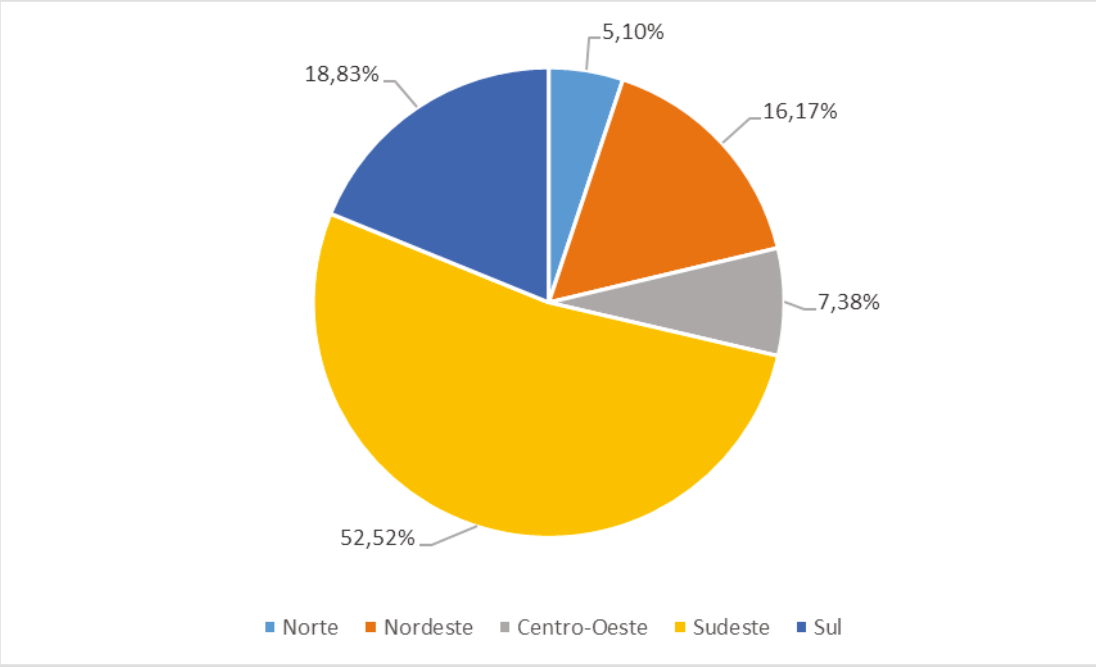
anteriormente, as pessoas com algum tipo de deficiência atingiam 44.073.377 pessoas em idade ativa em 2010, porém, 23,7 milhões não estavam ocupadas (IBGE, 2010).

Tendo todos os assuntos relacionados já tratados para uma melhor compreensão dos objetivos do presente trabalho, as seções a seguir são voltadas a alusão dos trâmites práticos da presente pesquisa, a luz do processo do KDD.

3 ESTATÍSTICA DESCRITIVA - BASE DE DADOS DO SUL

De acordo com o que fora explicitado anteriormente, a base CAGED refere-se cadastro dos empregados e desempregados do Brasil. Conforme será detalhado posteriormente na seção de aplicação (fase seleção dos dados), originalmente coletou-se 294.595.191 registros (cada registro indica uma admissão ou uma demissão), distribuídos entre os anos 2009 e 2016. Destes, retirou-se os registros com dados inconsistentes como, por exemplo, salário mensal = 0, o que resultou em 292.779.165 registros, dos quais as PcD representavam 99.53% enquanto as pessoas com deficiência representavam 0.47%, ou seja, 1.370.150 registros - que posteriormente foram selecionados. O Gráfico 8 a seguir evidencia a distribuição das regiões do Brasil (Centro-Oeste, Nordeste, Norte, Sudeste e Sul) nos dados selecionados até o momento:

GRÁFICO 8 - DISTRIBUIÇÃO DAS REGIÕES DO BRASIL NA BASE DE DADOS CAGED GERAL



FONTE: A AUTORA (2017).

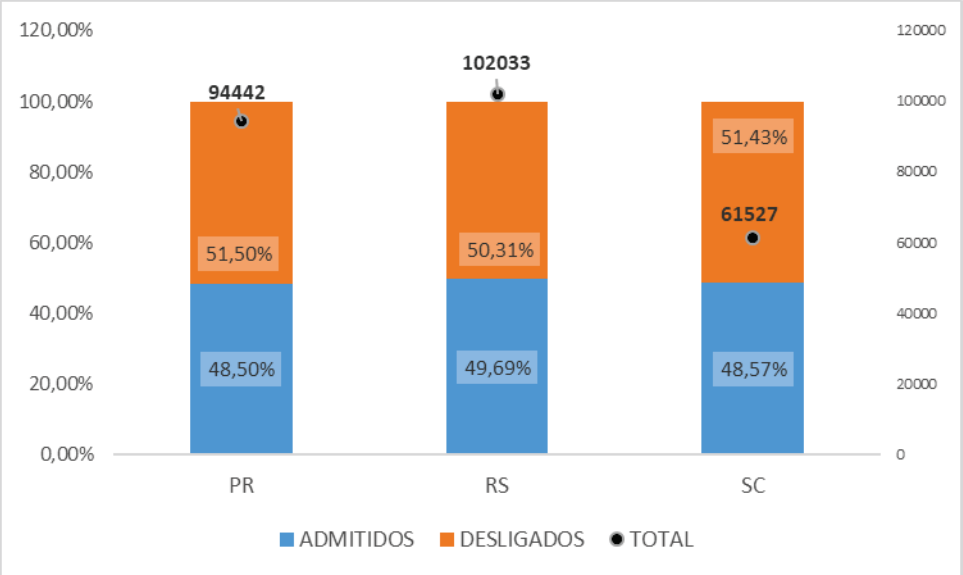
Conforme verifica-se no Gráfico 8, a região Norte é a menos representada nos registros de pessoas com deficiência (PcD), com apenas 5,10% (69.868) do total de 1.370.150 dados. Em seguida vem as regiões Centro-Oeste, Nordeste e Sul, com

representações de 7,38%; 16,17% e 18,83%, respectivamente. Destaca-se região Sudeste, cuja representação é de 52,52% dos dados, o que equivale a 719.517 registros.

Considerando as PcD, a região Sul representa 18,83% dos registros, quando comparado com o restante do Brasil, o que a consolida como a 2ª colocada quando se trata de proporção dos dados, pois, só perde para a região Sudeste, que possui 33,62% a mais dos dados totais. Por sua vez, o Sul possui 13,73% do total de dados a mais do que a região Norte (ou seja, 188.134 registros), 11,45% do total a mais do que a região Centro-Oeste (ou seja, 156.835 registros) e apenas 2,66% a mais do que a região Nordeste (ou seja, 36.441 registros). Para fins de mineração de dados, selecionou-se somente os registros de pessoas com deficiência e pertinentes à região Sul cuja sub-base compreende 258.002 registros. A escolha da região Sul foi devido à restrição de hardware e tempo.

Na região Sul 126.386 (48,99%) são admitidos e o restante, 131.616 (51,01%) demitidos, apresentados no Gráfico 9 a seguir por estado que compreende - Santa Catarina (SC), Paraná (PR) e Rio Grande do Sul (RS):

GRÁFICO 9 - DISTRIBUIÇÃO DE ADMITIDOS E DESLIGADOS POR ESTADO NA BASE DE DADOS DA REGIÃO SUL DO BRASIL

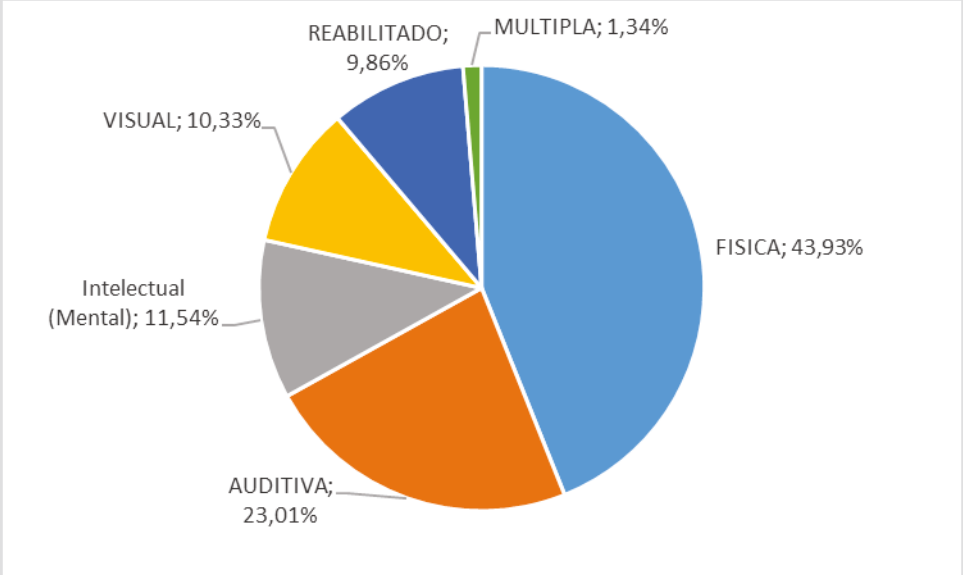


FONTE: A AUTORA (2017).

Pelo Gráfico 9 percebe-se que o estado do Rio Grande Sul se destaca quando se trata de quantidade de registros, compreendendo mais dados (102.033 que

equivalem à 39,55%), seguido do Paraná (94.442 registros = 36,61%) e de Santa Catarina (61.527, cuja representação é de 23,85%). Percebe-se também que os deficientes desligados são a maioria em todas UF da região Sul, com representação de 51,08% em média. A distribuição dos tipos de deficiência na base de dados selecionada é visualizada pelo Gráfico 10 a seguir:

GRÁFICO 10 - DISTRIBUIÇÃO DOS TIPOS DE DEFICIÊNCIA NA BASE DE DADOS DA REGIÃO SUL DO BRASIL

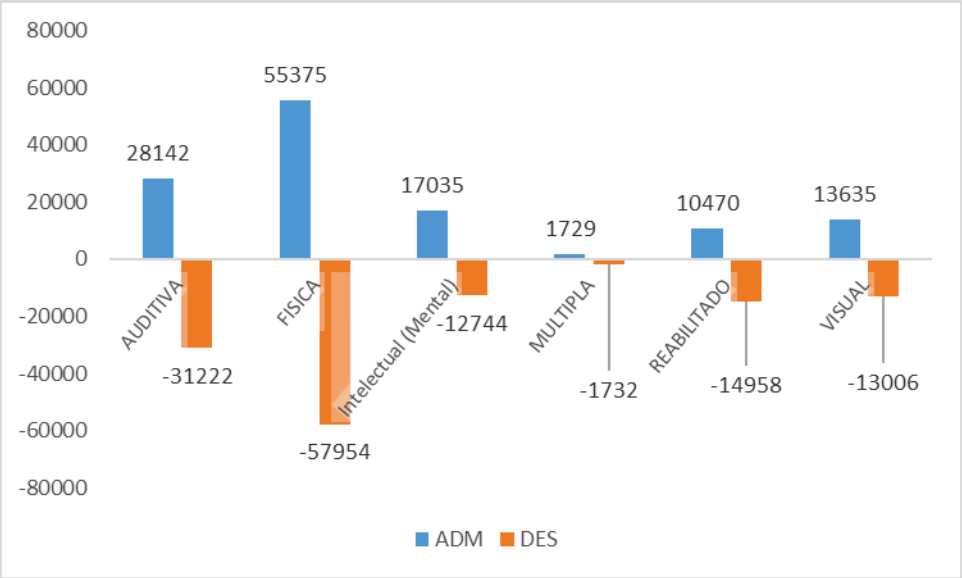


FONTE: A AUTORA (2017).

A partir do Gráfico 10 percebe-se que os deficientes físicos são a maioria, abrangendo 43,93% dos 258.002 registros da base do Sul. Em segundo lugar estão os deficientes auditivos, que representam 23,01% dos deficientes cadastrados na referida base. Posteriormente seguem os deficientes intelectuais, visuais, reabilitados e múltiplos, que correspondem a 11,54%; 10,33%; 9,68% e 1,34%, respectivamente.

De uma forma geral, o gráfico 11 a seguir indica o saldo de movimentação por deficiência –cujos valores negativo indicam desligamento e respectivamente, os valores positivos indicam admissão:

GRÁFICO 11 - SALDO DE MOVIMENTAÇÃO POR TIPO DE DEFICIÊNCIA NA BASE DE DADOS DA REGIÃO SUL DO BRASIL

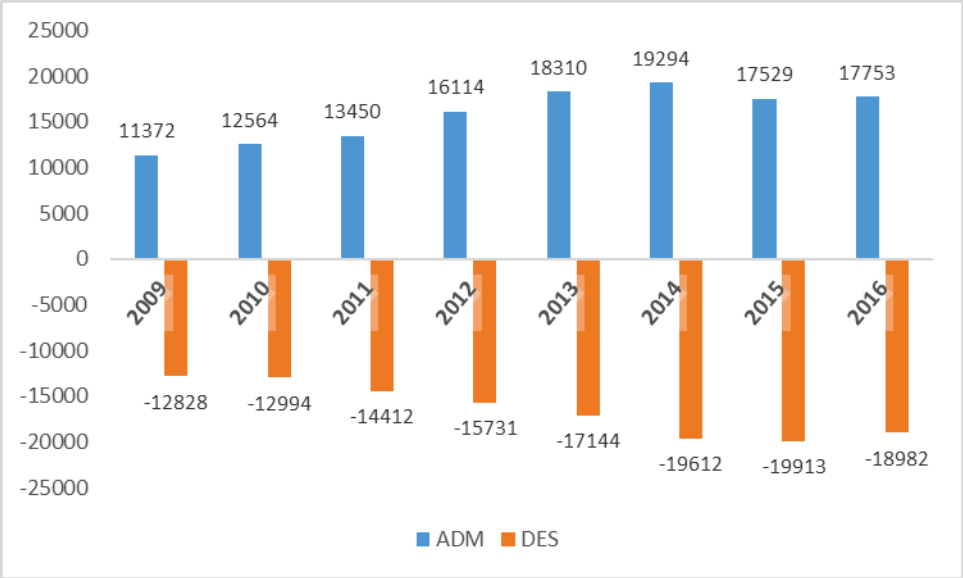


FONTE: A AUTORA (2017).

Identifica-se por intermédio do Gráfico 11 que o tipo de deficiência física se destaca por atingir o maior saldo nos dois tipos de movimentações: 55.375 pessoas admitidas e 57.954 pessoas desligadas. Por outro lado, estão os deficientes múltiplos, cujo saldo é o menor nos dois tipos de movimentações (1.729 deficientes contratados para 1.732 desligados). O resultado da soma dos admitidos e desligados (ignorando o sinal negativo) indica o total de registros por deficiência, onde destaca-se os deficientes físicos em quantidade de registros, fato já mostrado no Gráfico 3.

Além de demonstrar o saldo da movimentação por ano (2009 a 2016), cujos valores positivos indicam o saldo admissão e os valores negativos o saldo de desligamento, resultantes da soma do atributo *saldomov*, o Gráfico 12 também evidencia a distribuição dos dados dos deficientes por ano - pois, quando somados tais valores, (ignorando o sinal negativo) o resultado indica o número de registros por ano.

GRÁFICO 12 - SALDO DE MOVIMENTAÇÃO POR ANO NA BASE DE DADOS DA REGIÃO SUL DO BRASIL



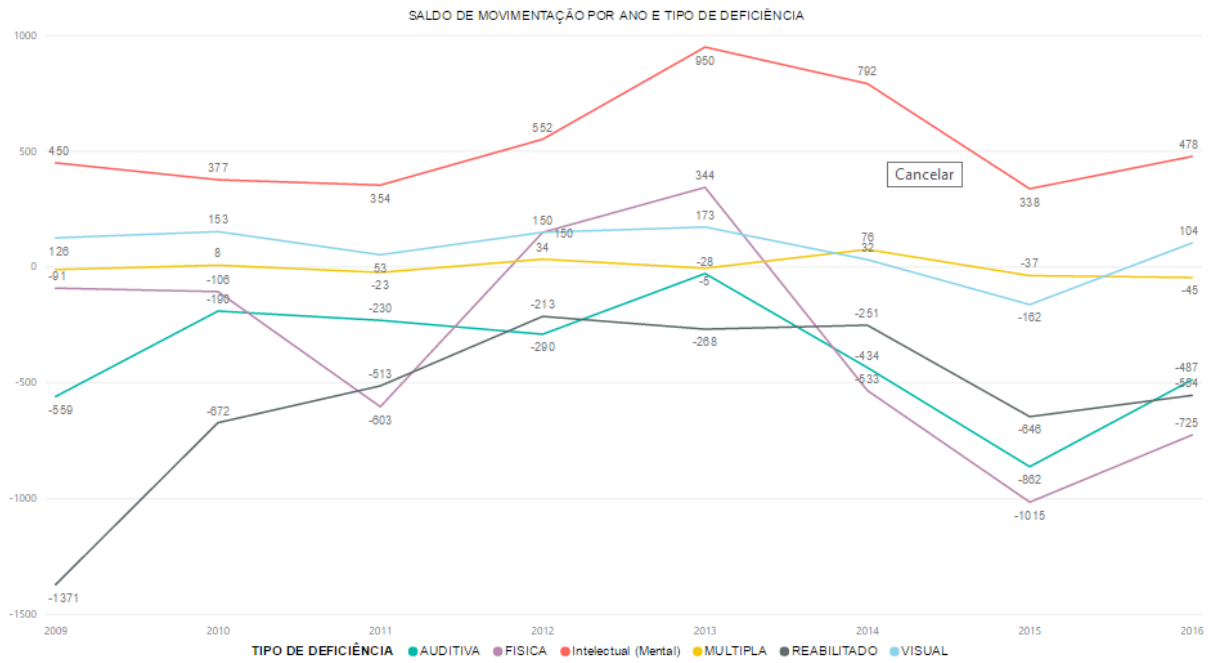
FONTE: A AUTORA (2017).

Analisando o saldo de movimentação pelo Gráfico 12, percebe-se que o ano de 2014 é o qual teve o maior saldo de admissão (19.294), cujo valor representa 15,27% do total de admitidos do Sul. Enquanto o saldo de desligamento fora maior no ano de 2015 (-19.913), cujo valor representa 15,13% do total de desligamentos da base. Em ordem decrescente e seguida do percentual de participação nos dados, o saldo de admissão se organiza da seguinte forma: 2013 (14,49%); 2016 (14,05%); 2015 (13,87%); 2012 (12,75%); 2011 (10,64%); 2010 (9,94%) e 2009 (9,00%). Já o menor saldo de desligamento encontra-se em 2009 (-12.828), cujo valor representa 9,75% do total de desligados, seguida de 2010 (9,87%), 2011 (10,95%), 2012 (11,95%), 2013 (13,03%), 2016 (14,42%), 2014 (14,90%) e finalmente 2015, já apresentado anteriormente.

Quanto ao número de registros (somatória dos valores por ano, ignorando o sinal negativo), percebe-se que 2014 é o ano que se destaca, com representação de 15,08% (38.906 registros), seguida dos anos 2015 (14,51%), 2016 (14,24%) e 2013 (13,74%), respectivamente. Nos últimos lugares em escala crescente, encontram-se os anos 2009, 2010, 2011 e 2012, com representações de 9,38%; 9,91%; 10,80% e 12,34% dos dados, respectivamente.

O gráfico 13 a seguir mostra a relação entre as variáveis tipo de deficiência e o ano, cujos valores resultam da soma do *saldomov*:

GRÁFICO 13 - SALDO DE MOVIMENTAÇÃO POR ANO E TIPO DE DEFICIÊNCIA NA BASE DE DADOS DA REGIÃO SUL DO BRASIL

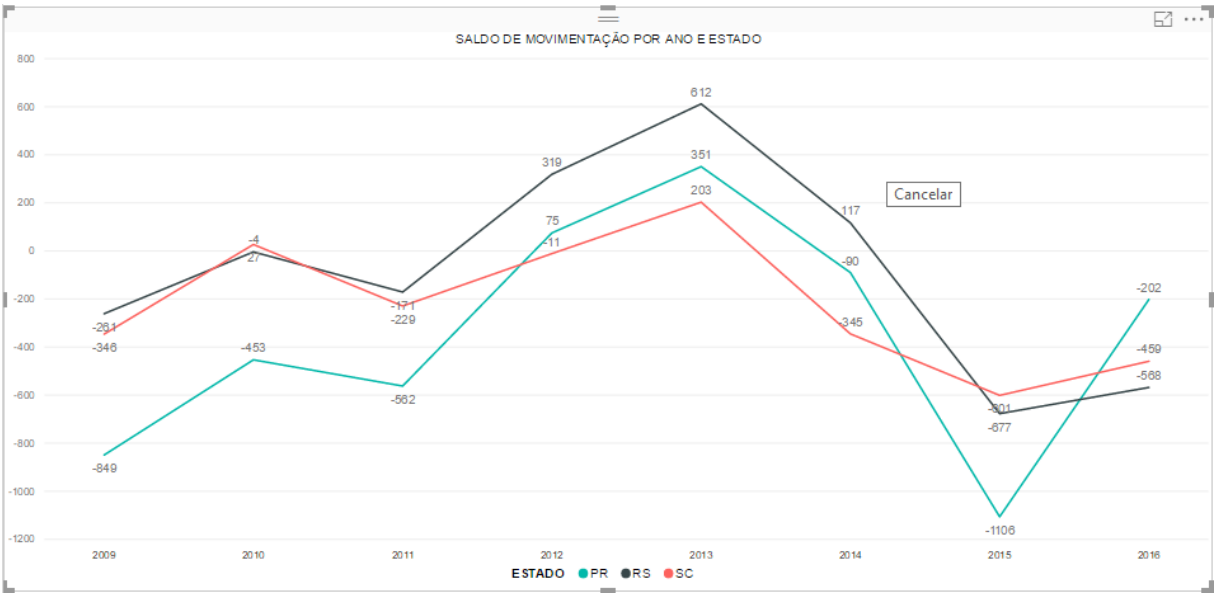


FONTE: A AUTORA (2017).

Por meio do Gráfico 13, percebe-se que ao longo dos anos (2009 e 2016) os deficientes físicos são os quais sofreram mudanças radicais (linha cor lilás), entre os saldos de movimentações, chegando a ser os mais demitidos duas vezes dentro do período: 2011 e 2015, com -603 e - 1015, respectivamente. Identifica-se um equilíbrio de admissão e demissão para os deficientes múltiplos (linha amarela) e visuais (linha azul). De uma forma geral, os deficientes intelectuais (linha vermelha) são os quais possuem um saldo sempre positivo e acima dos demais tipos de deficiência, atingindo o seu ápice em 2013, com 950 admitidos e em 2015 apenas 338 admitidos, que apesar de positivo, é o seu valor mínimo.

Já o Gráfico 14 a seguir, apresenta a relação entre as variáveis região e ano:

GRÁFICO 14 - SALDO DE MOVIMENTAÇÃO POR ANO E ESTADO DA REGIÃO SUL DO BRASIL

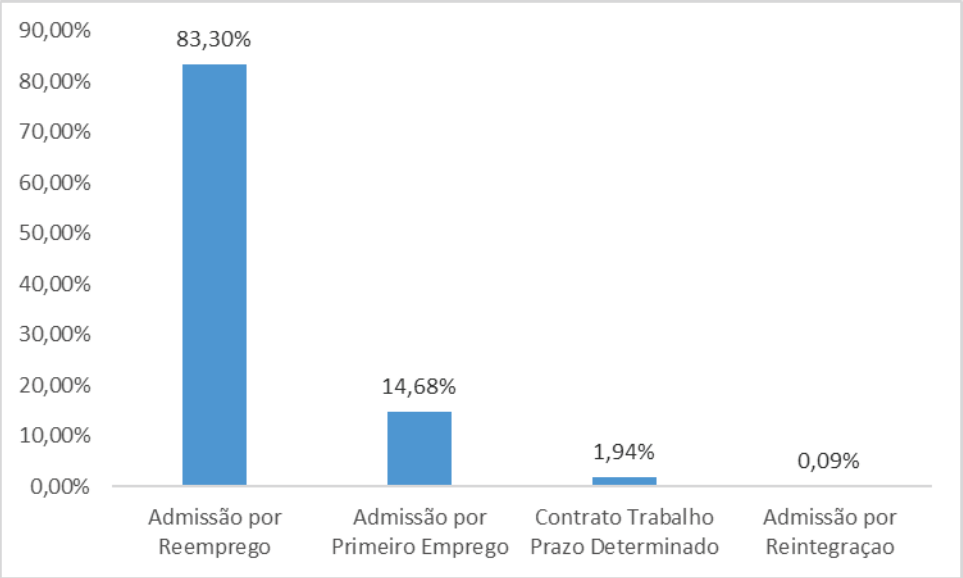


FONTE: A AUTORA (2017).

Segundo o Gráfico 14, destaca-se o Rio Grande do Sul (linha preta), com mais contratações no período de 2012 à 2014, com 319, 612 e 117, respectivamente. O Paraná (linha verde) possuiu apenas dois picos positivos: 2012, com 75 e 2013, com 351, o que destaca como o estado que possui o maior saldo de demissões em outros anos: -849 em 2009, -453 em 2010, -562 em 2011 e -1106 em 2015. Enquanto Santa Catarina (linha vermelha) possuiu o seu único saldo positivo em 2013 (203) e foi destaque em 2014 como o estado com o maior saldo negativo (-345).

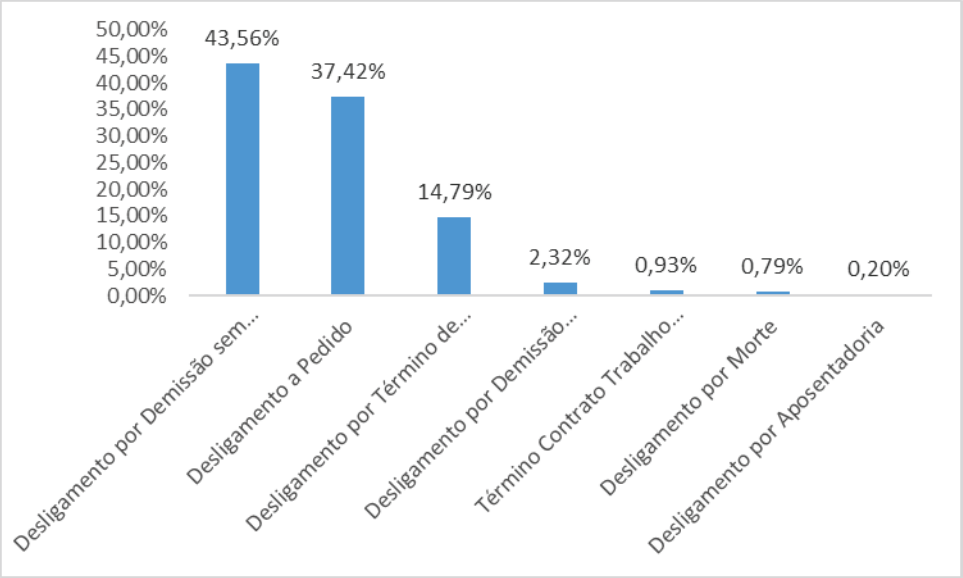
O atributo tipo de movimentação desagregada da base selecionada especifica o tipo de admissão e demissão dos cadastrados, contendo 11 variáveis distintas: Admissão por Primeiro Emprego, Admissão por Reemprego, Admissão por Reintegração, Contrato Trabalho Prazo Determinado, Desligamento a Pedido, Desligamento por Aposentadoria, Desligamento por Demissão com Justa Causa, Desligamento por Demissão sem Justa Causa, Desligamento por Morte, Desligamento por Término de Contrato e Término Contrato Trabalho Prazo Determinado, cuja distribuição - tipo de admissão e tipo de desligamento, é visualizada nos Gráficos 15 e 16, a seguir:

GRÁFICO 15 - DISTRIBUIÇÃO DOS TIPOS DE ADMISSÃO NA BASE DE DADOS DA REGIÃO SUL DO BRASIL



FONTE: A AUTORA (2017).

GRÁFICO 16 - DISTRIBUIÇÃO DOS TIPOS DE DESLIGAMENTO NA BASE DE DADOS DA REGIÃO SUL DO BRASIL



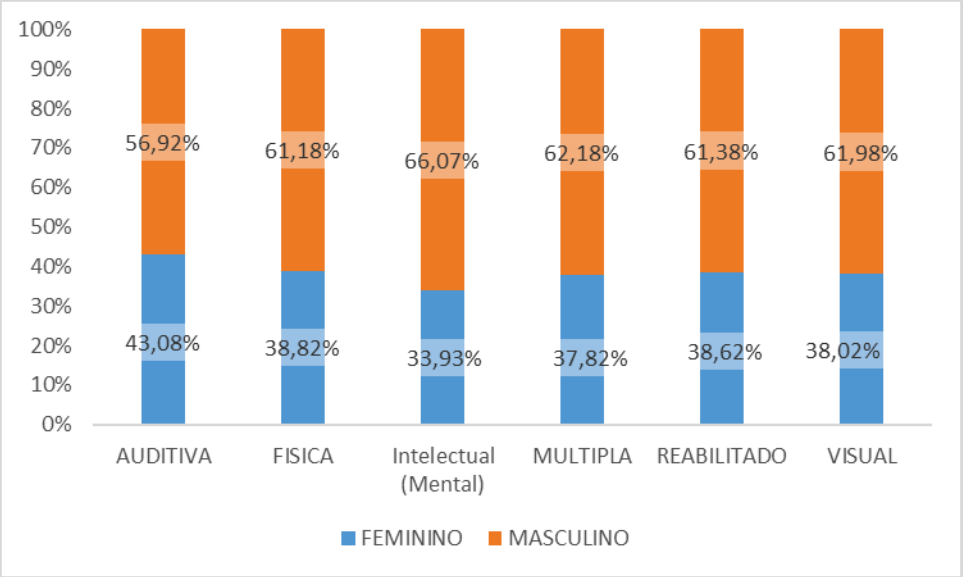
FONTE: A AUTORA (2017).

Conforme demostra o Gráfico 15, cujas variáveis já estão ordenadas de forma decrescente, Admissão por reemprego compreende 83,30% do total de admissões, enquanto no Gráfico 9, Desligamento por Demissão sem Justa Causa compreende 43,53% do total de desligamentos, seguida por Desligamento a Pedido com 37,42%.

Em ordem crescente e nos últimos lugares no Gráfico 15, estão os motivos Admissão por Reintegração (0,09%), Contrato Trabalho Prazo Determinado (1,94%) e Admissão por Primeiro Emprego (14,68%). Já no Gráfico 16, Desligamento por Aposentadoria (0,20%), Desligamento por Morte (0,79%), Término Contrato Trabalho Prazo Determinado (0,93%), Desligamento por Demissão com Justa Causa (2,32%), Desligamento por Término de Contrato (14,79%).

Quanto ao perfil de gênero, a referida base de dados compreende 157.073 homens (60,88%) e 100.929 mulheres (39,12%), apresentados no Gráfico 17 a seguir, por tipo de deficiência:

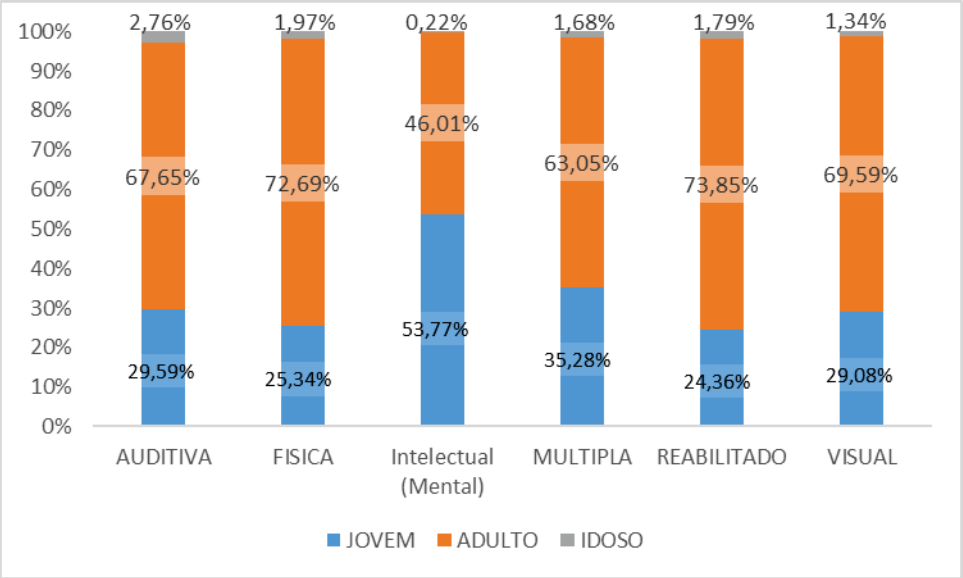
GRÁFICO 17 - DISTRIBUIÇÃO DO SEXO POR TIPO DE DEFICIÊNCIA NA BASE DE DADOS DA REGIÃO SUL DO BRASIL



FONTE: A AUTORA (2017).

Assim como na base em um contexto geral, percebe-se diante do Gráfico 17 que o gênero masculino é predominante em todos os tipos de deficiência, pois, em média as mulheres representam 38,38% dos dados de cada tipo de deficiência e os homens em média 61,62%. Já quanto a faixa etária, os adultos representam 68,12% (175.741) dos registros da base do Sul, seguida dos jovens (30,02%) e idosos (1,86%), padrão que se repete ao analisar detalhadamente por tipo de deficiência no Gráfico 18 a seguir, onde em média, os adultos representam 65,47%, os jovens 32,90% e os idosos 1,63%.

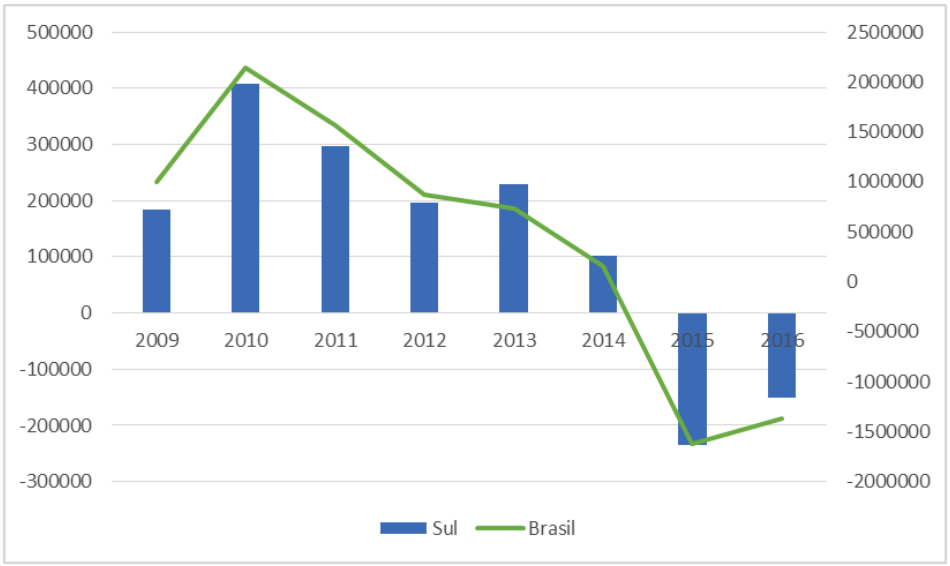
GRÁFICO 18 - DISTRIBUIÇÃO DA FAIXA ETÁRIA POR TIPO DE DEFICIÊNCIA NA BASE DE DADOS DA REGIÃO SUL DO BRASIL



FONTE: A AUTORA (2017).

Quando se trata do saldo de movimentação, identifica-se que o Sul seguiu a mesma tendência do Brasil dentro do período analisado, indicando que quando o ambiente econômico do país está favorável, isso consequentemente refletirá na região Sul com a geração de mais empregos, conforme indica o Gráfico 19:

GRÁFICO 19 - TENDÊNCIA SALDO DE MOVIMENTAÇÃO: UM COMPARATIVO ENTRE A REGIÃO SUL E O BRASIL



FONTE: A AUTORA (2017)

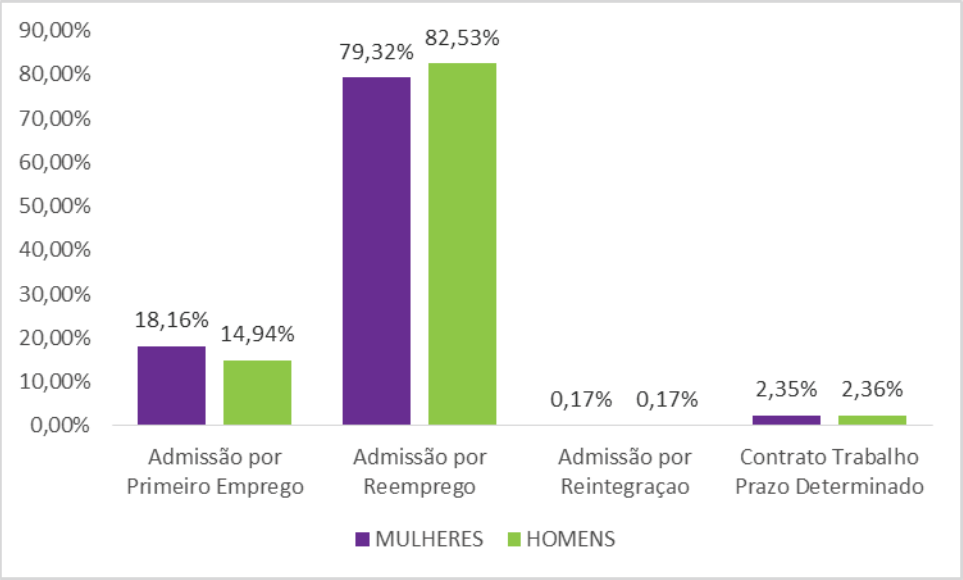
Porém, dentro do mesmo período, é evidente uma grande diferença do Sul com relação ao Brasil quando se trata do saldo de deficientes, conforme indica o Gráfico 20 a seguir, onde o Sul é representado pelas barras azuis e a linha verde o Brasil, indicando que o Sul sempre esteve acima da média do Brasil dentro do período analisado, exceto no ano 2015:

GRÁFICO 20 - SALDO DE DEFICIENTES: UM COMPARATIVO ENTRE A REGIÃO SUL E AS DEMAIS REGIÕES DO BRASIL

FONTE: A AUTORA (2017).

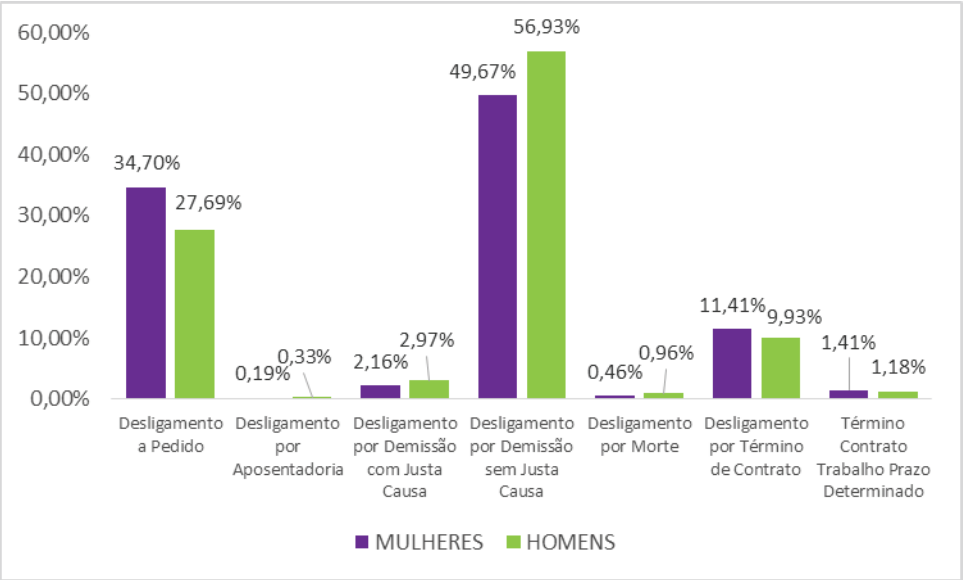
Relacionando o tipo de movimento desagregado com o sexo, é possível identificar qual o fator mais determinante para o desligamento e admissão entre os homens e mulheres. Os resultados apresentam-se nos Gráficos 21 e 22, sendo admissão e desligamento, respectivamente.

GRÁFICO 21 - DISTRIBUIÇÃO DOS TIPOS DE ADMISSÃO POR SEXO NA BASE DE DADOS DA REGIÃO SUL DO BRASIL



FONTE: A AUTORA (2017).

GRÁFICO 22 - DISTRIBUIÇÃO DOS TIPOS DE DESLIGAMENTO POR SEXO NA BASE DE DADOS DO SUL DO BRASIL



FONTE: A AUTORA (2017).

Os motivos predominantes nos registros que determinam o desligamento ou admissão por sexo é o mesmo para ambos os sexos. Sendo o destaque Admissão de reemprego no Gráfico 21, com representação de 79,32% entre os registros de admissão do sexo feminino e 82,53% nos registros de admissão somente do sexo

masculino. Já quando se trata de desligamento, Desligamento por demissão sem justa causa e a pedido novamente são destaques, entre ambos os sexos, com representação de 49,67% e 34,70%, respectivamente, nos registros de desligamentos somente do sexo feminino e 56,93% e 49,67%, respectivamente, somente nos registros de desligamentos do sexo masculino (Gráfico 22). Como é possível observar, não é diferente do contexto geral já apresentado (vide Gráficos 15 e 16) e os demais valores seguem até a mesma sequência.

Visto que a estatística descritiva teve por objetivo descrever e sumarizar os dados obtidos, na próxima seção detalha-se o processo prático de mineração de dados do presente trabalho.

4 APLICAÇÃO

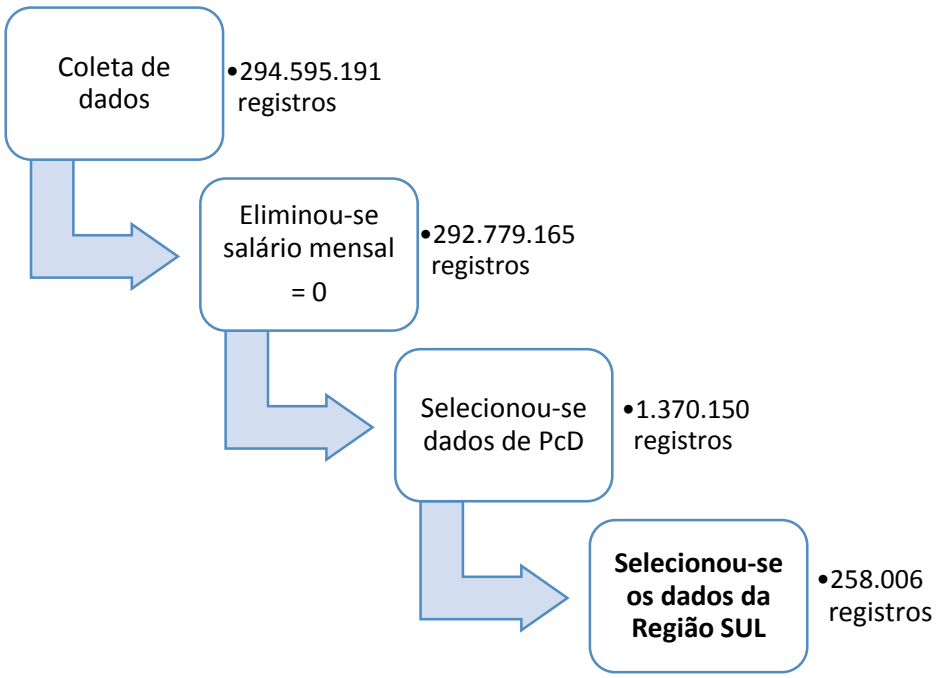
A presente aplicação conta com a exploração de uma grande base de dados, cuja tarefa da mineração de dados é realizar uma classificação com a finalidade de identificar padrões no mercado formal de trabalho das pessoas com deficiência da região Sul do Brasil. Conforme já explícito, o arquivo de dados a ser analisado contém registros de pessoas e estabelecimentos divulgados pelo CAGED - Cadastro Geral dos Empregados e Desempregados.

4.1 SELEÇÃO, PRÉ-PROCESSAMENTO E LIMPEZA DOS DADOS

Originalmente coletou-se 294.595.191 registros do período de 2009 à 2016 da base de dados abertos CAGED, onde cada registro indica uma admissão ou uma demissão. Posteriormente, tais registros foram armazenados em um banco de dados, visto que a ferramenta Excel não suportou o volume imenso de registros - cujo suporte máximo é de 1.048.576.

Após o armazenamento, eliminou-se os registros cujo valores no atributo salário mensal = 0, visto que eram dados inconsistentes. Diante disso, resultou em 292.779.165 registros, dos quais foram selecionados somente os registros de pessoas com deficiência (PcD), por meio de uma consulta SQL, onde eliminou-se os valores menores que 1 no atributo TipoDefic (tipo de deficiência), visto que os não deficientes tinham os seus respectivos registros = 0 no referido atributo e os demais valores representavam o tipo de deficiência que o dono de tal registro apresenta: 1 = Física; 2 = Auditiva; 3 = Visual; 4 = Intelectual (Mental); 5 = Múltipla; 6= Reabilitado. Diante disso, resultou em 1.370.150 registros de pessoas com deficiência (PcD). Finalmente, seletou-se os dados pertinentes à Região Sul (UF = PR, SC e RS), onde resultou em 258.006 registros, que posteriormente foram minerados. A Figura 19 a seguir indica as etapas de seleção dos dados, além das respectivas quantidades resultantes:

FIGURA 19: ETAPAS DE SELEÇÃO DOS DADOS



FONTE: A AUTORA (2017).

Tendo somente os registros que corroboram com o objetivo do presente trabalho, dos 40 atributos que a base traz consigo (vide APÊNDICE A), selecionou-se apenas os quais possuem relação direta com mesmos objetivos, restando 14 atributos a serem minerados: CBO2002Ocupacao; CNAE20Classe; FaixaEmprInicioJan; Grauinstrucao; QtdHoraContrat; Idade; IndAprendiz; RacaCor; SalarioMensal; Sexo; TempoEmprego; TipoDefic; TipoMovDesagregado e UF.

Sintetizando, para fins de mineração, extraiu-se da base de dados abertos CAGED 258.006 registros de pessoas com deficiência da região Sul do Brasil, distribuídos em 14 atributos.

4.2 TRANSFORMAÇÃO

Conforme apresentado no APÊNDICE A, originalmente os valores da base de dados CAGED são numéricos, onde cada valor assume um significado dentro da variável (vide Tabela 1). Diante disso, 10 das 14 variáveis tiveram os seus valores transformados para os reais significados e 4 atributos foram discretizados, visando a categorização de tais valores para facilitar a mineração de dados.

Os reais significados que substituíram os valores dos atributos transformados são os informados pelo arquivo *Layout*, disponibilizado no site do Ministério do Trabalho, já com a finalidade de explicar o significado de cada valor por atributo. O resultado da transformação pode ser consultado no ANEXO A.

Já a transformação dos atributos CBO2002Ocupacao e CNAE20Classe não ocorreu conforme determina do CAGED, visto que no arquivo Layout indicavam 2.584 e 676 valores distintos, respectivamente, o que tornaria complexo os resultados e dificultaria ainda mais a análise.

Os valores ou nomenclaturas da CBO 2002 são um conjunto de códigos e títulos que é utilizada na sua função enumerativa. É uma estrutura hierárquico-piramidal composta de:

a) 10 grandes grupos (GG)
b) 47 sete subgrupos principais (SGP)
c) 192 subgrupos (SG)
d) 596 grupos de base ou famílias ocupacionais (SG), onde se agrupam 2.422 ocupações e cerca de 7.258 títulos sinônimos.

Diante disso, a transformação dos valores apresentados originalmente no atributo CBO2002Ocupacao ocorreu para os grandes grupos, que é a categoria de classificação mais agregada, reunindo amplas áreas de emprego, mais do que tipos específicos de trabalho. Os grandes grupos de apresentam como o 1º número do código da família e são os seguintes, conforme indica a Tabela 1:

TABELA 1 - GRANDES GRUPOS - CBO 2002

(Continua)

1º Nº DO CÓDIGO DA FAMÍLIA	GRANDES GRUPOS/TÍTULOS	VALOR PARA MINERAÇÃO
0	Forças Armadas, Policiais e Bombeiros Militares	A
1	Membros superiores do poder público, dirigentes de organizações de interesse público e de empresas e gerentes	B
2	Profissionais das ciências e das artes	C
3	Técnicos de nível médio	D
4	Trabalhadores de serviços administrativos	E
5	Trabalhadores dos serviços, vendedores do comércio em lojas e mercados	F

(Continuação)

1º Nº DO CÓDIGO DA FAMÍLIA	GRANDES GRUPOS/TÍTULOS	VALOR PARA MINERAÇÃO
6	Trabalhadores agropecuários, florestais, da caça e pesca	G
7	Trabalhadores da produção de bens e serviços industriais	H
8	Trabalhadores da produção de bens e serviços industriais	
9	Trabalhadores de manutenção e reparação	I

FONTE: ADAPTADO DO MINISTÉRIO DO TRABALHO, 2017.

Devido a extensão dos títulos dos grandes grupos, os valores da CBO2002Ocupacao foram transformados para os elementos alfabéticos em sua ordem correspondente aos algarismos que representam os grandes grupos do respectivo registro, que por sua vez é de acordo com o primeiro elemento do seu código, resultando assim 8 valores distintos dentro do atributo (o que antes eram 2.584), visto que os grandes grupos 7 e 8 foram agrupados por possuírem o mesmo título, sendo representados por um único elemento alfabético (H).

O mesmo ocorreu com o atributo CNAE20Classe, optando pela classificação maior, denominada seções, visto que sua nomenclatura também é um conjunto de códigos e títulos que é utilizada na sua função enumerativa, sendo uma estrutura hierárquico-piramidal de:

a) 21 seções
b) 87 divisões
c) 285 grupos
d) 673 classes.
e) 1.301subclasses - não previsto no referido atributo, mas no atributo SB CLAS 20 descartado sim.

Diante disso, a transformação dos valores apresentados originalmente no atributo CNAE20Classe ocorreu para as seções, que é a categoria de classificação mais agregada, reunindo amplas classificações de atividades econômicas. As seções se apresentam em ordem alfabética e se apresentam como os dois primeiros número do código da família e são os seguintes, conforme indica a Tabela 2 a seguir:

TABELA 2 - SEÇÕES - CNAE 20 CLASSE

Seção	Divisões	Descrição CNAE
A	01 .. 03	AGRICULTURA, PECUÁRIA, PRODUÇÃO FLORESTAL, PESCA E AQÜICULTURA
B	05 .. 09	INDÚSTRIAS EXTRATIVAS
C	10 .. 33	INDÚSTRIAS DE TRANSFORMAÇÃO
D	35 .. 35	ELETRICIDADE E GÁS
E	36 .. 39	ÁGUA, ESGOTO, ATIVIDADES DE GESTÃO DE RESÍDUOS E DESCONTAMINAÇÃO
F	41 .. 43	CONSTRUÇÃO
G	45 .. 47	COMÉRCIO; REPARAÇÃO DE VEÍCULOS AUTOMOTORES E MOTOCICLETAS
H	49 .. 53	TRANSPORTE, ARMAZENAGEM E CORREIO
I	55 .. 56	ALOJAMENTO E ALIMENTAÇÃO
J	58 .. 63	INFORMAÇÃO E COMUNICAÇÃO
K	64 .. 66	ATIVIDADES FINANCEIRAS, DE SEGUROS E SERVIÇOS RELACIONADOS
L	68 .. 68	ATIVIDADES IMOBILIÁRIAS
M	69 .. 75	ATIVIDADES PROFISSIONAIS, CIENTÍFICAS E TÉCNICAS
N	77 .. 82	ATIVIDADES ADMINISTRATIVAS E SERVIÇOS COMPLEMENTARES
O	84 .. 84	ADMINISTRAÇÃO PÚBLICA, DEFESA E SEGURIDADE SOCIAL
P	85 .. 85	EDUCAÇÃO
Q	86 .. 88	SAÚDE HUMANA E SERVIÇOS SOCIAIS
R	90 .. 93	ARTES, CULTURA, ESPORTE E RECREAÇÃO
S	94 .. 96	OUTRAS ATIVIDADES DE SERVIÇOS
T	97 .. 97	SERVIÇOS DOMÉSTICOS
U	99 .. 99	ORGANISMOS INTERNACIONAIS E OUTRAS INSTITUIÇÕES EXTRATERRITORIAIS

FONTE: ADAPTADO DO IBGE, 2017.

Devido a extensão dos títulos, os valores do atributo CNAE20Classe foram transformados para os valores das seções apresentadas na tabela anterior, já instituído pelo próprio CNAE como elementos alfabéticos. Conforme apresenta a tabela anterior, a coluna divisões apresenta os dois primeiros itens do código que

correspondem aos elementos alfabéticos, resultando assim 21 valores (A-U) distintos dentro do atributo (o que antes eram 6767).

Os atributos QtdHoraContrat, Idade, SalarioMensal e TempoEmprego, cujos valores são inteiros ou contínuos e não possuem significados específicos impostos pelo próprio CAGED e nem tabelas específicas, foram discretizados, ou seja, tiveram os seus valores substituídos por intervalos a fim reduzir a quantidade final de valores, facilitando assim, o processo de mineração.

Castro e Ferrari (2016, p. 104) apresentam alguns métodos de discretização de dados:

a) Encaixotamento (*binning*) - onde ocorre a distribuição dos valores em intervalos (*bins*), cuja repartição é definida e substitui-se o valor de cada intervalo pela média ou mediana. Outra opção para substituição é a diferença entre o maior e menor valor do referido atributo, que será somado ao menor a fim de limitar o primeiro intervalo e assim sucessivamente, cuja aplicação fora selecionada será melhor detalhado na prática (vide Quadros 3, 4 e 5).
b) Análise de histograma - Similar ao encaixotamento, porém, são utilizadas as faixas do histograma para definir os intervalos de valores do atributo e posteriormente os valores do atributo são substituídos de acordo com a faixa na qual se encontram.
c) Agrupamento - Algoritmos de agrupamento são utilizados para particionar o atributo em grupos de valores seguindo um critério preestabelecido. Cada grupo de valores é representado por um protótipo que é utilizado em substituição aos valores que pertencem ao grupo.
d) Baseado na entropia - A *entropia* é uma medida baseada em informação que pode ser aplicada na segmentação dos valores de um atributo numérico *A,* dado um conjunto *D* de objetos.

Optou-se pelo método de encaixotamento para discretizar o atributo QtdHoraContrat, cujos detalhes apresentam-se na Quadro 2 seguir:

QUADRO 2 - PASSO A PASSO DISCRETIZAÇÃO ATRIBUTO QTDHORACONTRAT

PASSOS	**EXECUÇÃO**	
maior valor - menor valor	44 - 0 = 44	
resultado anterior/nº intervalos	44 / 3 = 14,67	
ESTRUTURAÇÃO		**RESULTADO**
intervalo 1	menor valor + resultado da divisão	0 + 14,67 = 14,67
intervalo 2	resultado anterior + resultado divisão	14,67 + 14,67 = 29,33
intervalo 3	resultado anterior + resultado divisão	29,33 + 14,76 = 44
ou seja,		
Jornada 1 | <= 14,67		
Jornada 2 | > 14,67 E < =29,33		
Jornada 3 | > 29,33		

FONTE: A AUTORA (2017).

Sintetizando, com base no seu menor valor (0) e maior valor (44), o atributo QtdHoraContrat fora discretizado em 3 intervalos, denominados Jornada 1, Jornada 2 e Jornada 3, que compreendem 1.882; 24.009 e 232.115 valores, respectivamente.

O mesmo método fora aplicado para a discretização do atributo TempoEmprego, cujos detalhes apresentam-se na Quadro 3 a seguir:

QUADRO 3 - PASSO A PASSO DISCRETIZAÇÃO ATRIBUTO TEMPOEMPREGO

PASSOS	**EXECUÇÃO**	
maior valor - menor valor	591 - 0 = 591	
resultado anterior/nº intervalos	591 / 3 = 197	
ESTRUTURAÇÃO		**RESULTADO**
intervalo 1	menor valor + resultado da divisão	0 + 197 = 197
intervalo 2	resultado anterior + resultado divisão	197 + 197 = 394
intervalo 3	resultado anterior + resultado divisão	394 +197 = 591
ou seja,		
Periodo 1 | <= 197		
Periodo 2 | > 197 E <= 394		
Periodo 3 | > 394		

FONTE: A AUTORA (2017).

Sintetizando, com base no seu menor valor (0) e maior valor (591), o atributo TempoEmprego fora discretizado em 3 intervalos, denominados Período 1, Período 2 e Período 3, que compreendem 253.422; 4.150 e 434 valores, respectivamente.

Aplicou-se o mesmo método para a discretização do atributo SalarioMensal, cujos detalhes apresentam-se na Quadro 4 a seguir:

QUADRO 4- PASSO A PASSO DISCRETIZAÇÃO ATRIBUTO SALARIOMENSAL

PASSOS	**EXECUÇÃO**	
maior valor - menor valor	119.415 - 0 = 119.415	
resultado anterior/nº intervalos	119.415 / 4 = 29.853,75	
ESTRUTURAÇÃO		**RESULTADO**
intervalo 1	menor valor + resultado da divisão	0 + 29.853,75 = 29.853,75
intervalo 2	resultado anterior + resultado divisão	29.853,75 + 29.853,75 = 59.707,50
intervalo 3	resultado anterior + resultado divisão	59.707,50 + 29.853,75 = 89.561,25
intervalo 4	resultado anterior + resultado divisão	89.561,25 + 29.853,75 = 119.415
ou seja,		
Faixa salarial 1 | <= 29.853,75		
Faixa salarial 2 | > 29.853,75 E <= 59.707,50		
Faixa salarial 3 | > 59.707,50 E <= 89.561,25		
Faixa salarial 4 | > 89.561,25		

FONTE: A AUTORA (2017).

Sintetizando, com base no seu menor valor (0) e maior valor (119.415), o atributo SalarioMensal fora discretizado em 4 intervalos, denominados Faixa salarial 1, Faixa salarial 2, Faixa salarial 3 e Faixa salarial 4, que compreendem 257.946; 48; 8 e 4 valores, respectivamente.

Já o atributo idade fora discretizado em três intervalos: jovens, adultos e idosos. O IBGE em suas estatísticas, denomina a faixa etária dos 15 aos 24 anos como a população jovem do Brasil, o que justificou tal seleção para a faixa denominada jovens. A Lei nº 10.741, popularmente conhecida como Estatuto do Idoso, regula os direitos das pessoas com idade igual ou superior a 60 anos, sendo este o critério para a faixa denominada idosos - o que logicamente levou a considerar os registros com idade entre 25 anos e 59 anos como adultos, ficando instituída as faixas conforme a Tabela 3:

TABELA 3 - DISCRETIZAÇÃO ATRIBUTO IDADE

INTERVALOS	VALORES
Jovem	> = 24 anos
Adulto	< = 25 anos E >= 59 anos
Idoso	>= 60 anos

FONTE: A AUTORA (2017).

4.3 ANÁLISE EXPLORATÓRIA E SELEÇÃO DE MODELOS

A presente etapa consiste na escolha da tarefa e algoritmo (s) para a efetiva mineração de dados, etapa subsequente. Como já dito, optou-se pela tarefa de classificação dos dados na ferramenta *WEKA*, cujos algoritmos disponíveis são apresentados na Tabela 4, a seguir:

TABELA 4 - LISTA DOS ALGORITMOS DE CLASSIFICAÇÃO ATIVOS NO *WEKA*

BAYES	RULES
BayesNet	DecisionTable
NaiveBayes	JRip
NaiveBayesMultinomialText	OneR
NaivesBayesSimple	PART
NaiveBayesUpdateable	ZeroR

META	TREES
AdaBoostM1	DecisionStump
AttributeSelectedClassifier	HoeffdingTree
Bagging	J48
ClassificationViaRegression	LMT
CVParameterSelection	RandomForest
FilteredClassifier	RandomTree
IterativeClassifierOptimizer	REPTree
LogitBoost	
MultiClassClassifier	
MultiClassClassifierUpdateable	
MultiScheme	
RandomCommittee	
RandomizableFilteredClassifier	
RandomSubSpace	
Stacking	
Vote	
WeightedInstancesHandlerwrapper	

FONTE: DADOS DA PESQUISA UTILIZANDO A FERRAMENTA *WEKA* (2017).

Diante disso, para fins de mineração de dados, considerou-se somente os algoritmos do tipo regras (*rules*) e árvores (*tree*): Além de serem os algoritmos clássicos que se baseiam na função classificação (SOARES JUNIOR e QUINTELLA, 2006, p. 1087), também são as heurísticas com melhor assimilação e/ou facilidade de entendimento dos resultados e segundo Berson at al (1999, apud PRASS, 2004, p. 21), a heurística de regras “é a melhor técnica de mineração de dados para expor todas as possibilidades de padrões existentes em um banco de dados.”

4.4 MINERAÇÃO DE DADOS – EXPERIMENTO 1

A referida etapa é a mineração de dados propriamente dita, onde inicialmente realizou-se dois experimentos, visando atingir dois objetivos distintos: O experimento A tendo o TipoMovDesagregado (Tipo do movimento desagregado) como atributo meta, a fim de buscar padrões que levam ao desligamento ou admissão dos deficientes da região Sul e o experimento B com o atributo TipoDefic (Tipo de Deficiência) como meta, a fim de conhecer o perfil dos deficientes do mercado formal de trabalho da região Sul do Brasil.

O atributo meta, também conhecido como atributo dependente, é o atributo selecionado cujo valor deseja testar (Castro e Ferrari, 2016, p. 70), ou seja, através dos valores do subconjunto dos demais atributos da base de dado, determina-se o valor do atributo meta (Silva, 2004, p. 8). Os valores de cada atributo meta selecionado, que deseja-se testar, podem ser consultados no ANEXO A.

A ferramenta *WEKA* disponibiliza recursos que possibilitam testar e validar os resultados, através de parâmetros de validade e confiabilidade nos modelos gerados (SILVA, 2004, p. 10), apresentados a seguir segundo Ferreira (2010) e Kirkby (2004 apud LIMA, 2005):

a) *Cross validation* - Método de validação em que os dados são divididos em N subconjuntos (blocos de dimensão semelhante - *folds*) para uma aprendizagem de N iterações. Ao longo do processo de treino são utilizados N - 1 blocos, e apenas um para teste, sendo este diferente a cada iteração. Este processo é repetido para as N amostras. A performance do classificador é definida de acordo com a média dos N testes. A vantagem da aplicação deste método prende-se, acima de tudo, com o fato de todos os dados serem utilizados.

b) *Suplied test set* - O teste pode ser feito também com um outro conjunto de dados como, o conjunto de teste, o qual poderá ser selecionado em Supplied test set através do botão Set, permitindo escolher, também, um outro arquivo para o teste. Então, se faz os testes com o conjunto clicando-se no botão Start (KIRKBY, 2004 apud LIMA, 2005).
c) *Use training set* - Segundo Kirkby (2004 apud Lima, 2005), é uma opção que usa, para realização de teste, o mesmo conjunto de treinamento utilizado para predição, que provem da seleção feita na aba anterior do Preprocess (Kirkby, 2004).
d) *Percentage split* - O conjunto de teste é escolhido de modo aleatório, habitualmente cerca de 20 a 30% dos elementos. Os restantes dados são alvo de treino e em seguida validados no conjunto reservado (conjunto de teste).

Exceto os modelos *LMT* e *RandomTree* do tipo *tree*, em ambos os experimentos todos os métodos das heurísticas regras e árvore de decisão foram testados em sua forma padrão (ou seja, seus respectivos parâmetros não foram modificados) na opção de teste validação cruzada (*cross validation*) também padrão (10 pastas). Os resultados apresentam-se na etapa seguinte por experimento realizado.

4.5 INTERPRETAÇÃO E AVALIAÇÃO

Antes de apresentar detalhadamente o resultado dos algoritmos selecionados, se faz necessário a apresentação dos indicadores estatísticos de avaliação de desempenho dos mesmos, visto que são outros recursos de validação dos resultados disponibilizado pela ferramenta *WEKA* e também são comum entre todos os diferentes métodos de classificação aplicado - apesar de cada um possuir suas peculiaridades de processamento dos dados e respectivos parâmetros que podem ser configurados.

As medidas de avaliação de desempenho de classificadores em geral trazem informações sobre algum tipo de taxa de acerto ou de erro do classificador para um ou mais conjuntos de dados, sendo esta a forma mais comum de avaliar o desempenho de um classificador, denominada *acurácia* (*preditiva*), segundo Castro e Ferrari (2016, p. 259). Como outros indicadores estão estatística Kappa, erro absoluto

médio, erro quadrado médio da raiz, erro absoluto relativo e erro relativo ao quadrado da raiz, bem como a precisão detalhada por classe de atributo meta e a matriz de confusão, cujos elementos são destaques no ANEXO B.

A seguir, detalha-se alguns dos indicadores estatístico (os mais populares) fornecido pelo *WEKA* à luz de Castro e Ferrari (2016) com complementação de outros autores (FERREIRA, 2010; MARTINEZ, CASAL e JANEIRO, 2009; SOCZEK e ORLOVSKI, 2014):

a) *Correctly Classified Instances: "*A taxa global de sucesso do algoritmo, conhecida como *acurácia* (*ACC*), é o número de classificações corretas dividido pelo número total de classificações" (CASTRO e FERRARI, 2016, p. 262).
b) *Incorrectly Classified Instances: É* o percentual de instâncias que o classificador previu incorretamente, sendo complementar *ao Correctly Classified Instances.*
c) *Kappa statistic: Segundo Castro e Ferrari (2016, p. 263), a estatística Kappa* mede a concordância de dois avaliadores, com cada um classificando *n* objetos em *C* classes mutuamente exclusivas [...]Se os avaliadores concordam plenamente, κ = 1; e se não há concordância, κ = 0. Corroborando e simplificando, Soczek e Orlovski (2014, p. 7) afirmam que tal indicador "mede o nível da concordância e ligação dos dados". Ferreira complementa que "um kappa igual a 1 indica concordância perfeita, enquanto um kappa igual a 0 indica concordância equivalente a um simples acaso." (FERREIRA, 2010, 67).
d) *Mean absolute error:* Segundo Martinez, Casal e Janeiro (2009), é a média da diferença entre os valores atuais e os preditos em todos os casos, é a média do erro da predição.
e) *Root mean squared error:* Segundo Martinez, Casal e Janeiro (2009), o referido indicador é "usado para medir o sucesso de uma predição numérica, [...] calculado pela média da raiz quadrada da diferença entre o valor calculado e o valor correto."
f) *Relative absolute error:* Segundo Martinez, Casal e Janeiro (2009), é o erro total absoluto. Em todas as mensurações de erro, valores mais baixos significam maior precisão do modelo, com o valor próximo de zero temos o modelo estatisticamente perfeito.
g) *Root relative squared error:* Segundo Martinez, Casal e Janeiro (2009), reduz o quadrado do erro relativo na mesma dimensão da quantidade sendo predita incluindo raiz quadrada. Assim como a raiz quadrada do erro significativo (root

mean-squared error), este exagera nos casos em que o erro da predição foi significativamente maior do que o erro significativo.

h) *TP Rate:* Taxa de verdadeiros positivos, ou seja, taxa de instâncias da classe positiva classificado como positivo (Castro e Ferrari, 2016, p. 260). Em outras palavras, Ferreira (2010, p. 64) corrobora com os autores: “é o número de previsões corretas para uma instância que é positiva”

i) *FP Rate:* Taxa de falsos positivos, ou seja, taxa de instâncias da classe negativa classificado como positivo (Castro e Ferrari, 2016, p. 260). Em outras palavras, Ferreira (2010, p. 64) corrobora com os autores: “é o número de previsões incorretas para uma instância que é negativa”. Martinez, Casal e Janeiro (2009) complementam que “são os dados classificados erroneamente como positivos pelo classificador.”

j) *Precision: “*Associadas ao conceito de *relevância*, a precisão mede a qualidade ou exatidão do algoritmo”, ou seja, “[...]a precisão mede a quantidade de objetos recuperados que são relevantes”. (Castro e Ferrari, 2016, p. 262). Já para Martinez, Casal e Janeiro (2009), o indicador “é o valor da predição positiva (número de casos positivos/total de casos cobertos), muito influenciada pela especificidade (número de casos negativos que são verdadeiramente negativos) e pouco pela sensitividade (número de casos positivos que são verdadeiramente positivos) [...]”.

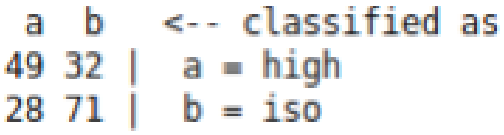
k) *Recall: “[...] A*ssociadas ao conceito de *relevância*, a revocação mede a completude do algoritmo”, ou seja, “[...]mede a quantidade de objetos relevantes que foram recuperados.” (Castro e Ferrari, 2016, p. 262). Já para Martinez, Casal e Janeiro (2009), “é o valor da cobertura de casos muito influenciada pela sensitividade e pouco pela especificidade. É calculada por número de casos cobertos pelo número total de casos aplicáveis.”

l) *F-Measure: C*onsiderando a precisão e a revocação, é usada para medir o desempenho, segundo Martinez, Casal e Janeiro (2009). Ferreira corrobora em outras palavras afirmando que “mede a eficácia de um classificador”.

m) *Confusion Matrix:* Segundo Ferreira (2010, p. 63), a matriz de confusão possibilita uma visualização evidente dos resultados de um determinado modelo: Os resultados são apresentados sob a forma de uma tabela de duas entradas (considerando problemas de apenas duas classes), sendo, uma delas constituída pelas classes desejadas, a outra pelas classes previstas pelo

modelo. As células, por sua vez, são preenchidas com o número de instâncias que correspondem ao cruzamento das entradas.” Vide a Figura 20 a seguir, exemplo dado pelo próprio autor:

FIGURA 20 - EXEMPLO DE MATRIZ DE CONFUSÃO.



FONTE: FERREIRA (2016, p. 64)

Para o exemplo acima, é possível constatar que no caso da classe *high*, de um universo de 81 instâncias, 49 foram classificadas corretamente e 32 incorretamente. Já no caso da classe *iso*, de um conjunto de 99 instâncias, 71 foram classificadas corretamente e 28 instâncias classificadas incorretamente. Diante disso, Castro e Ferrari (2016, p. 267) complementam que “[...] um bom classificador deve apresentar números maiores na diagonal principal e números pequenos, idealmente zero, fora dela.”

4.5.1 Experimento A1 - Padrão no mercado formal das PcD

Visando identificar padrões que levam ao desligamento/admissão no mercado formal das PcD na região Sul do Brasil, os resultados apresentados na Tabela 5 a seguir tiveram o atributo *TipoMovDesagregado* como meta, onde indica-se o desempenho de cada modelo por indicador de avaliação e simultaneamente destacam-se com cor amarelo os dois melhores desempenhos por indicador:

TABELA 5 - INDICADORES DE DESEMPENHO DOS RESULTADOS DO EXPERIMENTO A1 (PADRÃO NO MERCADO FORMAL DAS PCD)

		Correctly Classified Instances (%)	Kappa statistic	Mean absolute error	Root mean squared error	Relative absolute error (%)	Root relative squared error (%)	Tempo de processa-mento
RULES	*DecisionTable*	43.13	0.0857	0.1292	0.2533	96.46	97.90	88.8
	JRip	42.03	0.0281	0.1323	0.2573	98.77	99.41	156.07
	OneR	42.16	0.0285	0.1051	0.3243	78.51	125.31	24.92
	Part	41.99	0.1062	0.1252	0.2565	93.49	99.11	2977.14
	ZeroR	40.80	0	0.1339	0.2588	100	100	4.99
TREES	DecisionStump	41.27	0.0287	0.1324	0.2573	98.87	99.43	16.03
	HoeffdingTree	42.97	0.0919	0.1268	0.2541	94.69	98.19	216.02
	J48	44.10	0.1055	0.1265	0.2531	94.48	97.81	256.98
	RandomTree	36.41	0.0507	0.1286	0.2839	96.03	109.70	125.92
	REPTree	42.85	0.1007	0.1261	0.2551	94.14	98.60	6.47

FONTE: DADOS DA PESQUISA UTILIZANDO A FERRAMENTA *WEKA* (2017).

Os desempenhos destacados na Tabela 5, são melhores esclarecidos a seguir:

Quando comparados com outros algoritmos, é notório que *J48* (44,10%) e o *DecisionTable* (43,13%) apresentaram a maior taxa de classificações corretas (*Correctly Classified Instances)*, porém, tais taxas não são significativas, visto que ainda são inferior à 50%.

Quando se trata do nível de concordância indicado pela estatística Kappa (*Kappa statistic*), os algoritmos *Part* (0.1062) e *J48* (0.1055) se destacam por terem os respectivos valores mais próximos à 1 quando comparados com outros modelos, porém, tal resultado é inconcludente, visto que os valores não expressam uma correlação significativa.

Os algoritmos que apresentaram melhores desempenhos no Erro Médio Absoluto (*Mean Absolut Error*) foram o *OneR* (0.1051) e o *Part* (0.1252), respectivamente, visto que são os quais apresentam menores diferenças entre os valores atuais e preditos, quando comparados com os demais modelos.

Já para o indicador do Erro Quadrado Médio (*Root Mean Squared Error),* os modelos que se destacam são o *J48* (0.2531) e o *DecisionTable* (0.2533), visto que tais valores indicam menor erro entre os valores atuais e preditos, quando comparados com os demais algoritmos.

Os algoritmos *OneR* (78,51%) e *Part* (93,49%) são os quais se destacam no indicador de avaliação Erro Absoluto Relativo (*Relative Absolut Error*) por apresentarem os menores percentuais, indicando maior precisão dos respectivos modelo, quando comparado com os demais modelos.

Os destaques no indicador de avaliação Raiz Do Erro Quadrado Relativo (*Root Relative Squared Error*) são os modelos *J48* (97,81%) e *DecisionTable* (97,90%), respectivamente. Apesar dos percentuais ainda serem considerados altos, são os quais apresentam menor erro entre os valores atuais e preditos, quando comparados com os demais modelos.

Nota-se que não há uma relação entre o menor tempo de processamento e os melhores desempenhos indicados anteriormente, visto que neste aspecto destacam-se os modelos *ZeroR* (4.99 segundos) e *REPTree* (6.47 segundos), com os menores segundos de processamento, respectivamente, sendo estes não citados em nenhum indicador de avaliação como os melhores desempenhos.

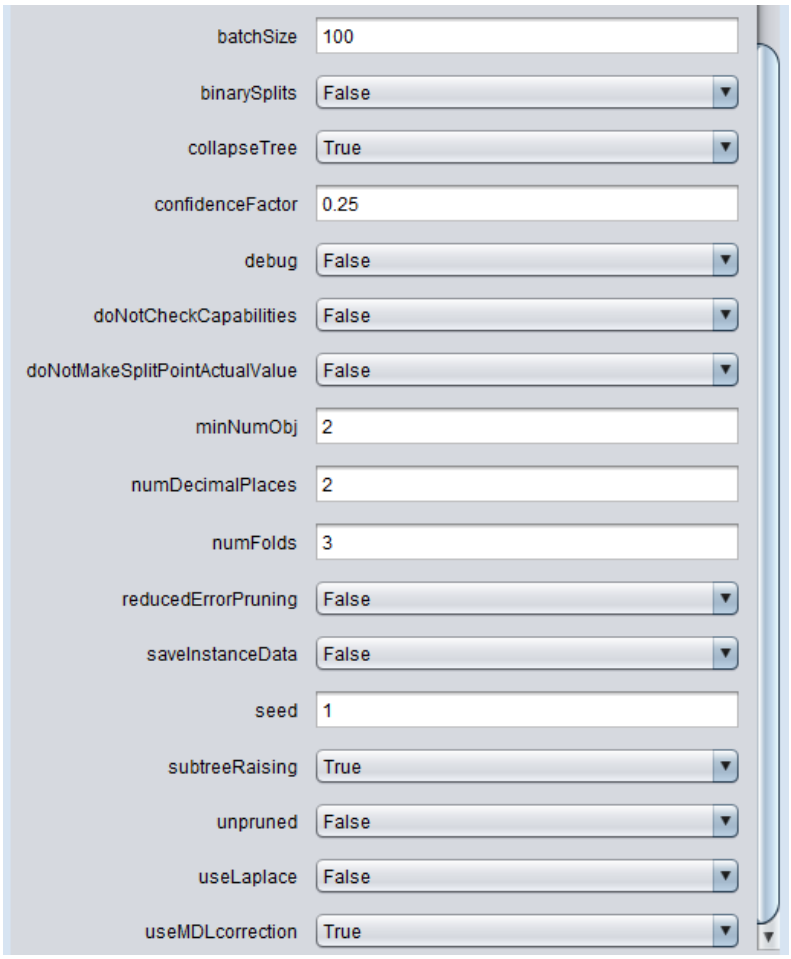
Diante disso, o modelo J48 é o primeiro selecionado para a demonstração detalhada dos respectivos resultados, visto que se destacou em 4 (quatro) dos 7 (sete) indicadores de avaliação de desempenho: classificação correta (*Correctly Classified Instances*), estatística Kappa (*Kappa statistic*), Erro Quadrado Médio (*Root mean squared error)* e Raiz do Erro Quadrado Relativo (*Root relative squared error).*

Já para o segundo selecionado houve um empate entre os modelos *DecisionTable* e *Part*, visto que ambos atingiram melhores desempenhos em 3 (três) dos 7 (sete) indicadores de avaliação de desempenho, sendo o DecisionTable nos seguintes indicadores: classificação correta (*Correctly Classified Instances*), Erro Quadrado Médio (*Root Mean Squared Error)* e Raiz do Erro Quadrado Relativo (*Root Relative Squared Error) e o Part* nos seguintes: estatística Kappa (*Kappa Statistic*), Erro Absoluto Médio (*Mean Absolute Error)* e Erro Absoluto Relativo *(Relative absolute Error).* Visto que o *Part* atingiu melhor desempenho na estatística Kappa, apresentando melhor concordância (apesar de tal valor não ser significativo, conforme apresentado anteriormente) e seus resultados são facilmente compreensíveis, o que levou-se a considerá-lo como o segundo modelo para o detalhamento dos resultados, indicados nas subseções a seguir.

4.5.1.1 J48

A ferramenta *WEKA* define o algoritmo J48 como “classe para gerar uma árvore de decisão C4.5 podada ou não” (tradução nossa). Visto ser equivalente ao C.4.5, Castro e Ferrari (2016, p. 283) complementam que o C.4.5 é uma versão aperfeiçoada do ID3, o algoritmo básico para indução de árvores de decisão, proposto por J. R. Quinlan. Os parâmetros do J48, em sua forma padrão, se apresentam conforme a Figura 21 no *WEKA*:

FIGURA 21 - PARÂMETROS J48 NA FERRAMENTA *WEKA*



FONTE: *WEKA* (2017).

A seguir detalha-se cada parâmetro, conforme indicado na própria ferramenta (tradução nossa) e relacionados na mesma ordem em que se apresenta:

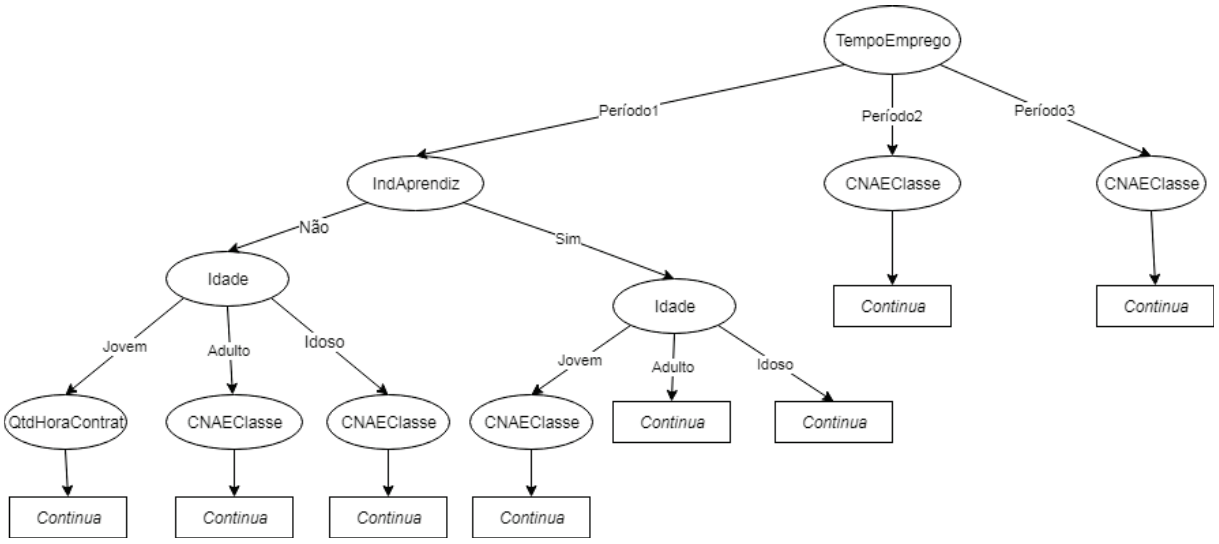
a) *BatchSize* - O número preferido de instâncias para processar se a predição do lote está sendo executada. Mais ou menos instâncias podem ser fornecidas,

mas isso dá às implementações a chance de especificar um tamanho de lote preferido.

b) *BinarySplits* - Se deseja usar divisões binárias em atributos nominais ao construir as árvores.
c) *CollapseTree* – Para remover as peças que não reduzem o erro de treino.
d) *ConfidenceFactor*- O fator de confiança usado para a poda (os valores menores incorrem em mais podas).
e) *Debug* - Se definido como *true*, o classificador pode gerar informações adicionais no console.
f) *DoNotCheckCapabilities* - Se configurado, os recursos do classificador não são verificados antes que o classificador seja construído (Use com cuidado para reduzir o tempo de execução).
g) *DoNotMakeSplitPointActualValue* - Se verdadeiro, o ponto de divisão não é realocado para um valor de dados real. Isso pode gerar acelerações substanciais para grandes conjuntos de dados com atributos numéricos.
h) *MinNumObj* - O número mínimo de instâncias por folha.
i) *NumDecimalPlaces* - O número de casas decimais a serem usadas para a saída de números no modelo.
j) *NumFolds* - Determina a quantidade de dados utilizados para a poda de erros reduzidos. Uma dobra é usada para a poda, o resto para o crescimento da árvore.
k) *ReducedErrorPruning* - Se a poda de erro reduzido é usada em vez da poda do C.4.5.
l) *SaveInstanceData* - Se deseja salvar os dados de treinamento para visualização.
m) *Seed*- A semente usada para randomizar os dados quando é utilizada a poda de erro reduzido.
n) *SubtreeRaising* - Se considerar a operação de levantamento da subarrânea na poda.
o) *Unpruned* - Se a poda é realizada.
p) *UseLaplace* - Se contagem das folhas é alisado com base em Laplace.
q) *UseMDLcorrection* - Se a correção MDL é usada ao encontrar divisões em atributos numéricos.

Com os seus parâmetros na sua forma padrão e tendo como atributo meta o *TipoMovDesagregado para o* presente experimento, o algoritmo J48 apresentou como resultado uma árvore de decisão com 4.286 folhas e tamanho 5.056. Apesar do seu extenso tamanho, é possível perceber pelo resultado gráfico gerado que a raiz da árvore é o atributo TempoEmprego, onde se desdobram os primeiros nós, representados por IndAprendiz (para o TempoEmprego = período1) e o CNAEClasse (para o TempoEmprego = período2 e período 3). No terceiro nível identifica-se o atributo idade, tanto para o IndAprediz sim ou não, porém para o CNAEClasse resultante dos períodos 2 e 3 não foi possível verificar os próximos nós ou folhas devido a sobreposição dos mesmos. No quarto nível destacam-se os atributos HoraContrat e o CNAEClasse, porém, a partir do quinto nível não é possível verificar os próximos nós ou folhas devido a sobreposição dos mesmos. Os atributos identificados estão conforme indica a Figura 22 a seguir:

FIGURA 22 - IDENTIFICAÇÃO DA RAÍZ E DOS PRIMEIROS NÍVEIS/NÓS DA ÁRVORE DE DECISÃO DO J48 NO EXPERIMENTO A1 (PADRÃO NO MERCADO FORMAL DAS PCD)



FONTE: A AUTORA COM BASE NOS RESULTADOS DA PESQUISA UTILIZANDO A FERRAMENTA *WEKA* (2017).

Os atributos identificados no topo da árvore de decisão resultante da aplicação do J48 no presente experimento que visa identificar padrões de desligamento/admissão no mercado formal de PcD da região Sul do Brasil, são destaques por se constituírem os fatores que mais se correlacionam com o atributo

meta selecionado (TipoMovDesagregado), ou seja, são os fatores importantes que levam ao padrão de desligamento ou admissão dos deficientes na região Sul do Brasil. A sobreposição dos nós e folhas geradas na continuação da árvore, resultado do seu tamanho extenso, impossibilitou uma análise completa, cuja árvore original pode ser consultada no Anexo C.

Visando explorar tais resultados sob uma outra perspectiva, a Figura 23 a seguir apresenta detalhadamente os valores pertinentes à acurácia de tal experimento:

FIGURA 23 - ACURÁCIA J48 - EXPERIMENTO A1 (PADRÃO NO MERCADO FORMAL DAS PCD)

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0,927	0,822	0,437	0,927	0,594	0,150	0,575	0,455	Admissão por Reemprego
0,184	0,048	0,520	0,184	0,271	0,209	0,661	0,390	Desligamento por Demissão sem Justa Causa
0,160	0,017	0,422	0,160	0,232	0,227	0,754	0,227	Admissão por Primeiro Emprego
0,012	0,007	0,297	0,012	0,023	0,024	0,575	0,228	Desligamento a Pedido
0,000	0,000	0,000	0,000	0,000	-0,001	0,617	0,020	Desligamento por Demissão com Justa Causa
0,081	0,011	0,380	0,081	0,133	0,147	0,685	0,199	Desligamento por Término de Contrato
0,164	0,003	0,385	0,164	0,230	0,247	0,758	0,169	Contrato Trabalho Prazo Determinado
0,000	0,000	0,000	0,000	0,000	-0,000	0,675	0,020	Desligamento por Morte
0,078	0,000	0,606	0,078	0,138	0,217	0,684	0,052	Desligamento por Aposentadoria
0,101	0,001	0,383	0,101	0,160	0,195	0,771	0,130	Término Contrato Trabalho Prazo Determinado
0,000	0,000	0,000	0,000	0,000	0,000	0,660	0,003	Admissão por Reintegraçao
0,441	0,350	0,416	0,441	0,337	0,143	0,619	0,350	

FONTE: DADOS DA PESQUISA UTILIZANDO A FERRAMENTA *WEKA* (2017).

Conforme é possível verificar na Figura 23, os valores obtidos no *True Positive* (*TP Rate*) indicam uma probabilidade mais alta de um verdadeiro positivo na proporção de 92% na classe Admissão por reemprego, seguido pela classe Desligamento por demissão sem justa causa (18%), o que nos possibilita identificar que que o padrão detectado acerta mais vezes nos casos de Admissão por reemprego e Desligamento por demissão sem justa causa, respectivamente, do que em Desligamento por morte e Admissão por reintegração, que juntos apresentaram o menor TP representado por 0%.

Conforme já apresentado, os resultados obtidos em *False Positive* (*FP Rate*) indicam os dados classificados erroneamente como positivos, sendo a maior probabilidade encontrada em Admissão por reemprego com 82% e a menor em Desligamento por demissão por justa causa, por morte, por aposentadoria e por reintegração, que juntos possuem 0%.

Como dito anteriormente, a precisão indica o valor da predição positiva (número de casos positivos por total de casos cobertos). Os resultados foram melhores para a classificação de Desligamento por aposentadoria, apresentando 60%

de precisão, enquanto Desligamento por demissão por justa causa, por morte e por reintegração apresentaram 0%.

Visto que corresponde ao valor da cobertura de casos, o resultado do *recall* (cobertura) é equivalente aos valores obtidos no verdadeiro positivo (TP).

Os valores obtidos no *F-Measure* indicam um desempenho de 59% para a classificação de Admissão por reemprego e de 0% para a classificação de Desligamento por demissão por justa causa, por morte e por reintegração.

Os resultados obtidos em MCC demonstram um desempenho de 24% para a classificação de Contrato trabalho por prazo determinado. O *ROC Área* do classificador Término contrato trabalho por tempo determinado ficou em 77%, representando um bom resultado quando comparado com as demais classes, visto que valores próximos a 1 definem um classificador como ótimo. A pior classificação representada por 57% Admissão por reemprego e Desligamento a Pedido. Por fim, o *PRC Área* do classificador Admissão por reemprego ficou em 45%, enquanto a pior classificação foi Admissão por reintegração com 0,0%.

A matriz de confusão do referido algoritmo no presente experimento é indicada na Figura 24.

FIGURA 24 - MATRIZ DE CONFUSÃO J48 - EXPERIMENTO A1 (PADRÃO NO MERCADO FORMAL DAS PCD)

```
    a     b     c     d     e     f     g     h     i     j     k   <-- classified as
97591  4890  1197   512     4   951   100     5     0    29     0 |     a = Admissão por Reemprego
45830 10531   374   393     3   118    76     4     4     3     0 |     b = Desligamento por Demissão sem Justa Causa
14004   409  2970   224     2   898    31     0     0    10     0 |     c = Admissão por Primeiro Emprego
44276  3080   865   583     0   252   148     2     5    37     0 |     d = Desligamento a Pedido
 2696   269    25    41     0    11     6     0     0     2     0 |     e = Desligamento por Demissão com Justa Causa
15789   478  1404    82     1  1568   122     0     0    16     0 |     f = Desligamento por Término de Contrato
 1375    97   130    66     0   278   403     0     0   101     0 |     g = Contrato Trabalho Prazo Determinado
  625   362    10    37     0     7     1     0     4     0     0 |     h = Desligamento por Morte
  131    93     1     9     0     2     0     1    20     0     0 |     i = Desligamento por Aposentadoria
  792    25    60    15     0    45   160     0     0   123     0 |     j = Término Contrato Trabalho Prazo Determinado
   97    10     1     4     0     0     0     0     0     0     0 |     k = Admissão por Reintegraçao
```

FONTE: DADOS DA PESQUISA UTILIZANDO A FERRAMENTA *WEKA* (2017).

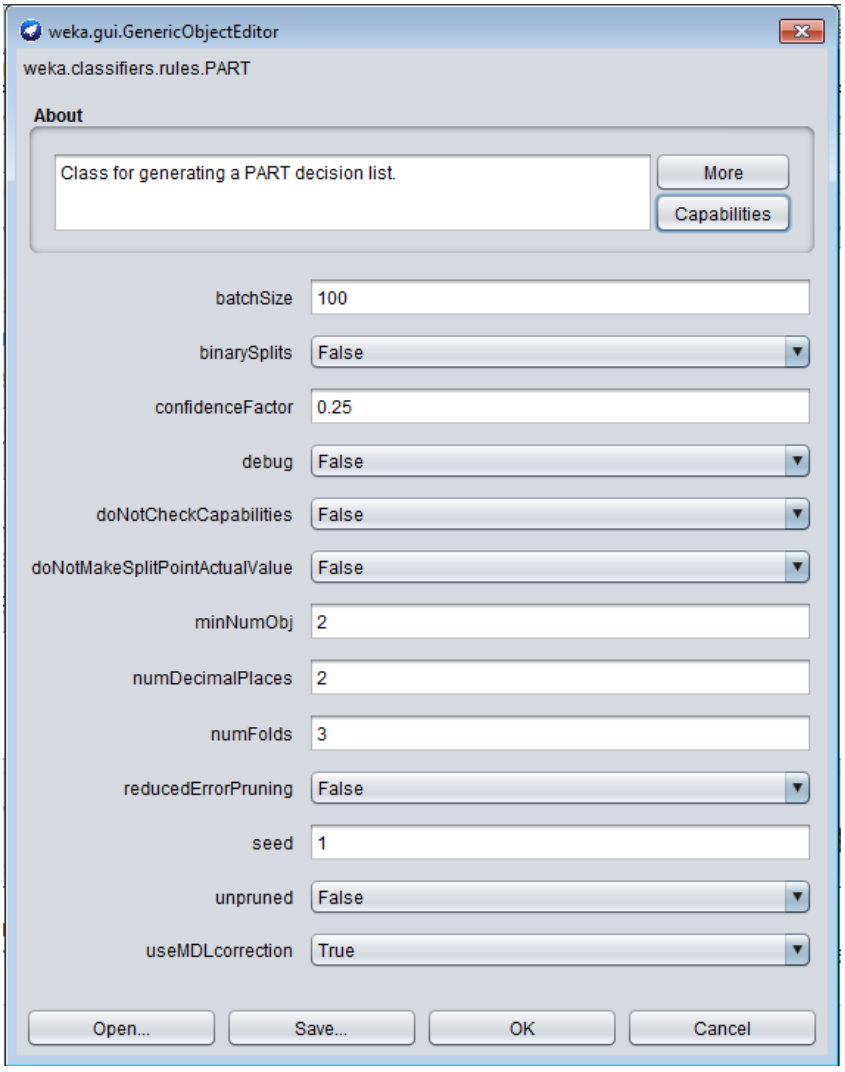
A matriz de confusão da Figura 24 possibilita outro ângulo dos mesmos resultados, visto que os valores contidos na diagonal principal é a quantidade de registros classificada corretamente por classe, conforme explicitado anteriormente (vide Figura 20), é possível identificar que a classe Desligamento por demissão com justa causa, Desligamento por morte e Admissão por reintegração não obtiveram nenhum registro classificado corretamente (representados por 0 na diagonal

principal). Por outro lado, possibilita observar que a classe Admissão por reemprego é a qual obteve mais classificações corretas, o que 93% dos seus registros (97.591 na diagonal principal).

4.4.1.2 Part

O algoritmo *Part* é definido pelo *WEKA* como “Algoritmo que gera uma lista de decisão. Usa a estratégia de divisão e conquista. Constrói de forma parcial uma árvore de decisão C4.5 em cada iteração, transformando a “melhor folha” numa regra.” (Tradução nossa). Seus parâmetros, em sua forma padrão, são indicados na Figura 25 a seguir:

FIGURA 25 - PARÂMETROS PART NA FERRAMENTA *WEKA*



FONTE: *WEKA* (2017).

Notoriamente os parâmetros são os mesmos do algoritmo J48, apresentado anteriormente, visto que ambos são derivados do C.4.5. A única exceção é o J48 conter o parâmetro *CollapseTree* a mais. Diante disso, dispensa-se tais detalhes para o algoritmo Part.

No referido experimento, o algoritmo Part apresentou 10.046 regras, das quais destacam-se 10 na Tabela 6 a seguir, cujo critério de seleção foi o número de casos corretamente cobertos pela respectiva regra maior ou igual a 600 (indicado pela terceira e última coluna da tabela), devido ao volume de registros da base processada.

TABELA 6 - 10 REGRAS DESTACADAS DO ALGORITMO PART - EXPERIMENTO A1

(Continua)

	REGRA - ALGORITMO PART	**INTERPRETAÇÃO**	**Casos cobertos corretamente/casos cobertos incorretamente**
1	TempoEmprego = periodo 2 AND CNAE20Classe = C AND Idade = adulto AND UF = RS: Desligamento por Demissão sem Justa Causa	*TempoEmprego = entre 197 e 394 meses E CNAE20Classe = Indústrias de transformação E Idade = entre 25 e 59 anos E UF = RS, ENTÃO: Desligamento por Demissão sem Justa Causa*	985.0/98.0
2	IndAprendiz = não AND Idade = jovem AND QtdHoraContrat = jornada 3 AND CNAE20Classe = C AND TipoDefic = REABILITADO AND UF = SC AND FaixaEmprInicioJan = 1000 OU MAIS: Admissão por Reemprego	*IndAprendiz = não E Idade = até ou igual à 24 anos E QtdHoraContrat = acima de 29h33min semanais E CNAE20Classe = Indústrias de transformação E TipoDefic = reabilitado E UF = SC E FaixaEmprInicioJan = 1000 ou mais, ENTÃO: Admissão por Reemprego*	907.24/389.8
3	IndAprendiz = não AND Idade = adulto AND CNAE20Classe = G AND QtdHoraContrat = jornada 3 AND CBO2002Ocupacao = D AND GrauInstrucao = Médio Completo AND FaixaEmprInicioJan = DE 10 A 19: Admissão por Reemprego	*IndAprendiz = não E Idade = entre 25 e 59 anos E CNAE20Classe = Comércio; reparação de veículos automotores e motocicletas E QtdHoraContrat = acima de 29h33min semanais E CBO2002Ocupacao = Técnicos de nível médio E GrauInstrucao = Médio Completo E FaixaEmprInicioJan = de 10 a 19, ENTÃO: Admissão por Reemprego*	889.6/445.22

(Continuação)

	REGRA - ALGORITMO PART	**INTERPRETAÇÃO**	**Casos cobertos corretamente/casos cobertos incorretamente**
4	IndAprendiz = não AND Idade = adulto AND CNAE20Classe = C AND UF = SC AND CBO2002Ocupacao = G AND FaixaEmprInicioJan = 1000 OU MAIS AND GrauInstrucao = Médio Completo AND RacaCor = BRANCA: Admissão por Reemprego	*IndAprendiz = não E Idade = entre 25 e 59 anos E CNAE20Classe = Indústrias de transformação E UF = SC E CBO2002Ocupacao = Trabalhadores agropecuários, florestais, da caça e pesca E FaixaEmprInicioJan = 1000 ou mais E GrauInstrucao = Médio Completo E RacaCor = branca, ENTÃO: Admissão por Reemprego*	881.66/509.99
5	IndAprendiz = não AND Idade = jovem AND QtdHoraContrat = jornada 3 AND CNAE20Classe = C AND TipoDefic = FISICA AND GrauInstrucao = Médio Completo AND FaixaEmprInicioJan = 1000 OU MAIS AND CBO2002Ocupacao = G: Admissão por Reemprego	*IndAprendiz = não E Idade = até ou igual à 24 anos E QtdHoraContrat = acima de 29h33min semanais E CNAE20Classe = Indústrias de transformação E TipoDefic = física E GrauInstrucao = Médio Completo E FaixaEmprInicioJan = 1000 ou mais E CBO2002Ocupacao = Trabalhadores agropecuários, florestais, da caça e pesca, ENTÃO: Admissão por Reemprego*	751.32/403.66
6	IndAprendiz = não AND Idade = adulto AND CNAE20Classe = C AND CBO2002Ocupacao = G AND FaixaEmprInicioJan = 1000 OU MAIS AND TipoDefic = FISICA AND GrauInstrucao = 6ª a 9ª Fundamental: Admissão por Reemprego	*IndAprendiz = não E Idade = entre 25 e 59 anos E CNAE20Classe = Indústrias de transformação E CBO2002Ocupacao = Trabalhadores agropecuários, florestais, da caça e pesca E FaixaEmprInicioJan = 1000 ou mais E TipoDefic = física E GrauInstrucao = 6ª a 9ª Fundamental, ENTÃO: Admissão por Reemprego*	738.31/466.92
7	IndAprendiz = não AND Idade = adulto AND CNAE20Classe = C AND UF = PR AND CBO2002Ocupacao = G AND GrauInstrucao = Médio Completo AND TipoDefic = FISICA AND FaixaEmprInicioJan = 1000 OU MAIS AND RacaCor = BRANCA AND sexo = mas: Desligamento por Demissão sem Justa Causa	*IndAprendiz = não E Idade = entre 25 e 59 anos E CNAE20Classe = Indústrias de transformação E UF = PR E CBO2002Ocupacao = Trabalhadores agropecuários, florestais, da caça e pesca E GrauInstrucao = Médio Completo E TipoDefic = física E FaixaEmprInicioJan = 1000 ou mais E RacaCor = branca E sexo = masculino, ENTÃO: Desligamento por Demissão sem Justa Causa*	737.69/413.98

(Continuação)

REGRA - ALGORITMO PART	INTERPRETAÇÃO	Casos cobertos corretamente/casos cobertos incorretamente
IndAprendiz = não AND Idade = adulto AND CNAE20Classe = G AND QtdHoraContrat = jornada 3 AND CBO2002Ocupacao = D AND TipoDefic = FISICA AND GrauInstrucao = Médio Completo AND FaixaEmprInicioJan = DE 20 A 49: Admissão por Reemprego	*IndAprendiz = não E Idade = entre 25 e 59 anos E CNAE20Classe = Comércio; reparação de veículos automotores e motocicletas E QtdHoraContrat = acima de 29h33min semanais E CBO2002Ocupacao = Técnicos de nível médio E TipoDefic = física E GrauInstrucao = Médio Completo E FaixaEmprInicioJan = de 20 a 49, ENTÃO: Admissão por Reemprego*	699.74/360.85
IndAprendiz = não AND Idade = adulto AND CNAE20Classe = H AND CBO2002Ocupacao = E AND TipoDefic = FISICA AND UF = PR: Admissão por Reemprego	*IndAprendiz = não E Idade = entre 25 e 59 anos E CNAE20Classe = Transporte, armazenagem e correio E CBO2002Ocupacao = Trabalhadores de serviços administrativos E TipoDefic = física E UF = PR, ENTÃO: Admissão por Reemprego*	626.82/349.02
IndAprendiz = não AND Idade = adulto AND CNAE20Classe = C AND CBO2002Ocupacao = G AND FaixaEmprInicioJan = DE 250 A 499 AND UF = RS AND RacaCor = BRANCA: Admissão por Reemprego	*IndAprendiz = não E Idade = entre 25 e 59 anos E CNAE20Classe = Indústrias de transformação E CBO2002Ocupacao = Trabalhadores agropecuários, florestais, da caça e pesca E FaixaEmprInicioJan = de 250 a 499 E UF = RS E RacaCor = branca, ENTÃO: Admissão por Reemprego*	624.64/382.63

FONTE: DADOS DA PESQUISA UTILIZANDO A FERRAMENTA *WEKA* (2017).

Devido a quantidade de atributos selecionados para tal experimento, é evidente que as regras apresentadas são bem específicas, conforme indica a coluna Interpretação. A única regra selecionada que contempla o Paraná é a 9, onde os deficientes físicos que não são aprendiz, possuem idade entre 25 e 59 anos, trabalham no setor de Transporte, armazenagem e correio e ainda ocupam cargos

relacionados a serviços administrativos, a tendência é que sejam admitidos por Reemprego.

Visando explorar tais resultados sobre outra perspectiva, a Figura 26 a seguir apresenta detalhadamente os valores pertinentes à acurácia de tal experimento:

FIGURA 26 - ACURÁCIA PART- EXPERIMENTO A1 (PADRÃO NO MERCADO FORMAL DAS PCD)

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0,248	0,093	0,432	0,248	0,315	0,193	0,668	0,386	Desligamento por Demissão sem Justa Causa
0,812	0,721	0,437	0,812	0,568	0,105	0,563	0,444	Admissão por Reemprego
0,051	0,045	0,213	0,051	0,083	0,013	0,569	0,222	Desligamento a Pedido
0,089	0,014	0,335	0,089	0,141	0,141	0,695	0,208	Desligamento por Término de Contrato
0,193	0,003	0,373	0,193	0,255	0,264	0,777	0,192	Contrato Trabalho Prazo Determinado
0,200	0,029	0,348	0,200	0,254	0,222	0,757	0,237	Admissão por Primeiro Emprego
0,000	0,000	0,000	0,000	0,000	-0,002	0,613	0,019	Desligamento por Demissão com Justa Causa
0,005	0,000	0,128	0,005	0,009	0,024	0,616	0,013	Desligamento por Morte
0,035	0,000	0,474	0,035	0,065	0,129	0,687	0,031	Desligamento por Aposentadoria
0,000	0,000	0,000	0,000	0,000	-0,000	0,550	0,005	Admissão por Reintegraçao
0,126	0,001	0,367	0,126	0,188	0,213	0,780	0,150	Término Contrato Trabalho Prazo Determinado
0,420	0,327	0,372	0,420	0,350	0,118	0,615	0,345	

FONTE: DADOS DA PESQUISA UTILIZANDO A FERRAMENTA *WEKA* (2017).

Conforme é possível notar na Figura 26, os valores obtidos no *True Positive* (*TP Rate*) indicam uma probabilidade mais alta de um verdadeiro positivo na proporção de 81% na classe Admissão por reemprego, seguido pela classe Desligamento por demissão sem justa causa (24%), o que nos possibilita identificar que que as regras detectadas acertam mais vezes nos casos de Admissão por reemprego e Desligamento por demissão sem justa causa, respectivamente, do que em Desligamento demissão com justa causa e Admissão por reintegração, que juntos apresentaram o menor TP representado por 0%.

Conforme já apresentado, os resultados obtidos em *False Positive* (*FP Rate*) indicam os dados classificados erroneamente como positivos, sendo a maior probabilidade encontrada em Admissão por reemprego com 72% e a menor em Desligamento por demissão por justa causa, por morte, por aposentadoria e por reintegração, que juntos possuem 0%.

Como dito anteriormente, a precisão indica o valor da predição positiva (número de casos positivos por total de casos cobertos). Os resultados foram melhores para a classificação de Desligamento por aposentadoria, apresentando 47% de precisão, enquanto Desligamento por demissão por justa causa e por reintegração apresentaram 0%.

Visto que que corresponde ao valor da cobertura de casos, o resultado do *recall* (cobertura) é equivalente aos valores obtidos no verdadeiro positivo (TP).

Os valores obtidos no *F-Measure* indicam um desempenho de 56% para a classificação de Admissão por reemprego e de 0% para a classificação de Desligamento por demissão com justa causa e por reintegração.

Os resultados obtidos em MCC demonstram um desempenho de 26% para a classificação de Contrato trabalho prazo determinado. O *ROC Área* do classificador Término contrato trabalho por tempo determinado ficou em 78%, representando um bom resultado quando comparado com as demais classes, visto que valores próximos a 1 definem um classificador como ótimo. A pior classificação representada por 56% Admissão por reemprego e Admissão por integração. Por fim, o *PRC Área* do classificador Admissão por reemprego ficou em 44%, enquanto a pior classificação foi Admissão por reintegração com 0,0%.

A matriz de confusão do referido algoritmo no presente experimento é indicada na Figura 27 a seguir:

FIGURA 27 - MATRIZ DE CONFUSÃO PART - EXPERIMENTO A1 (PADRÃO NO MERCADO FORMAL DAS PCD)

```
    a     b     c     d     e     f     g     h     i     j     k   <-- classified as
14208 39990  1927   207    92   869    17    13     2     0    11 |     a = Desligamento por Demissão sem Justa Causa
10540 85507  5142  1476   140  2389    16     9     0     1    59 |     b = Admissão por Reemprego
 4915 39460  2535   355   167  1738     8     9     6     0    55 |     c = Desligamento a Pedido
  881 14514   473  1737   160  1681     2     1     0     0    11 |     d = Desligamento por Término de Contrato
  231  1096   134   273   473   127     0     1     1     0   114 |     e = Contrato Trabalho Prazo Determinado
 1096 11254  1369  1042    57  3710     5     1     1     0    13 |     f = Admissão por Primeiro Emprego
  419  2351   170    18     8    81     0     0     0     0     3 |     g = Desligamento por Demissão com Justa Causa
  392   537    82    13     2    14     1     5     0     0     0 |     h = Desligamento por Morte
   98   129    16     2     1     1     1     0     9     0     0 |     i = Desligamento por Aposentadoria
   24    77     8     0     1     2     0     0     0     0     0 |     j = Admissão por Reintegraçao
   77   679    27    67   166    50     0     0     0     0   154 |     k = Término Contrato Trabalho Prazo Determinado
```

FONTE: DADOS DA PESQUISA UTILIZANDO A FERRAMENTA *WEKA* (2017).

A matriz de confusão da Figura 27 possibilita outro ângulo dos mesmos resultados, visto que os valores contidos na diagonal principal é a quantidade de registros classificada corretamente por classe, conforme já explicitado (vide Figura 20). É possível identificar que a classe Desligamento por demissão por justa causa e Admissão por reintegração não obtiveram nenhum registro classificado corretamente (representados por 0 na diagonal principal). Por outro lado, possibilita observar que a classe Admissão por reemprego é a classe que mais obteve classificações corretas, cuja representação é 81% dos seus registros (97.591 na diagonal principal).

4.5.2 Experimento B1 - Perfil das PcD

Visando o perfil das PcD na região Sul do Brasil, a presente aplicação teve o atributo *TipoDefic* como meta, a fim de identificar o perfil dos deficientes na Região Sul do Brasil. Os resultados apresentam-se na Tabela 7 a seguir, onde indica-se o desempenho de cada modelo por indicador de avaliação e simultaneamente destaca-se de amarelo os dois melhores desempenhos por indicador:

TABELA 7 - INDICADORES DE DESEMPENHO DOS RESULTADOS DO EXPERIMENTO B1 (PERFIL DAS PCD)

		Correctly Classified Instances (%)	Kappa statistic	Mean absolute error	Root mean squared error	Relative absolute error (%)	Root relative squared error (%)	Tempo de processa-mento
RULES	*DecisionTable*	*49.21*	*0.219*	*0.2211*	*0.3288*	*92.09*	*94.90*	*547.01*
	JRip	*47.32*	*0.0934*	*0.23*	*0.3392*	*95.79*	*97.91*	*1335.08*
	OneR	*45.44*	*0.0481*	*0.1819*	*0.4265*	*75.75*	*123.09*	*0.47*
	Part	*50.48*	*0.2453*	*0.2034*	*0.3286*	*84.73*	*94.83*	*2209.39*
	ZeroR	43.92	0	0.2401	0.3465	100	100	0.1
TREES	*DecisionStump*	*44.54*	*0.0615*	*0.2367*	*0.3441*	*98.61*	*99.30*	*0.89*
	HoeffdingTree	*49.42*	*0.2011*	*0.2091*	*0.3306*	*87.09*	*95.40*	*7.5*
	J48	*51.60*	*0.2416*	*0.208*	*0.3264*	*86.64*	*94.20*	*9.17*
	RandomTree	*47.92*	*0.2268*	*0.2004*	*0.3515*	*83.48*	*101.44*	*2.97*
	REPTree	*50.34*	*0.2288*	*0.2079*	*0.329*	*86.60*	*94.94*	*16.8*

FONTE: DADOS DA PESQUISA UTILIZANDO A FERRAMENTA *WEKA* (2017).

Os desempenhos destacados na Tabela 7, são melhores esclarecidos a seguir:

É notório que *J48* (51,60%) e o *Part* (50,48%) apresentaram a maior taxa de classificações corretas (*Correctly Classified Instances)* quando comparados com os demais modelos, sendo tais taxas significativas, visto que são superiores à 50%.

Quando se trata do nível de concordância indicado pela estatística Kappa (*Kappa statistic*), novamente os algoritmos *Part* (0.2453) e *J48* (0.2416) são destaques

por terem os respectivos valores mais próximos à 1 quando comparados com outros modelos, porém, tal resultado é também é inconcludente, visto que os valores não expressam uma correlação significativa.

Os algoritmos que apresentaram melhores desempenhos no Erro Médio Absoluto (*Mean Absolut Error*) foram o *OneR* (0.1819) e o *RandomTree* (0.2004), respectivamente, visto que são os quais apresentam menores diferenças entre os valores atuais e preditos, quando comparados com os demais modelos.

Já para o indicador do Erro Quadrado Médio (*Root Mean Squared Error),* os modelos que se destacam são o *J48* (0.3264) e o *Part* (0.3286) novamente, visto que tais valores indicam menor erro entre os valores atuais e preditos, quando comparados com os demais algoritmos.

Os algoritmos *OneR* (75,75%) e *DecisionTable* (83,48%) são os quais se destacam no indicador de avaliação Erro Absoluto Relativo (*Relative Absolut Error*) por apresentarem os menores percentuais, indicando maior precisão dos respectivos modelo, quando comparado com os outros modelos.

Os destaques no indicador de avaliação Raiz Do Erro Quadrado Relativo (*Root Relative Squared Error*) são os modelos *J48* (94,20%) e *Part* (94,83%), respectivamente. Apesar dos percentuais ainda serem considerados altos, são os quais apresentam menor erro entre os valores atuais e preditos, quando comparados com os demais modelos.

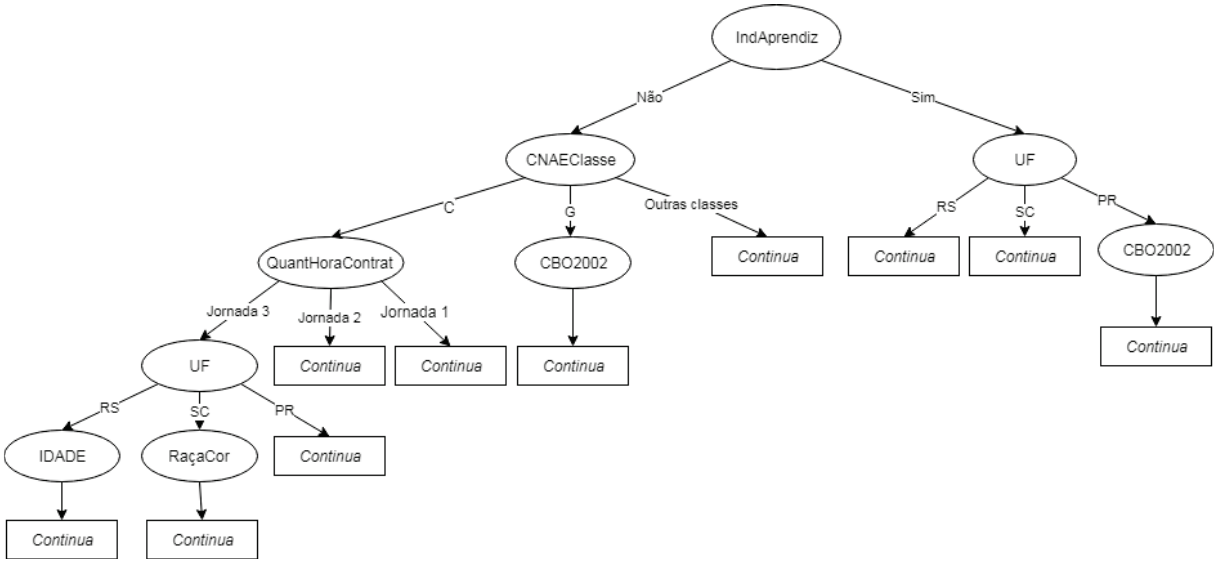
Nota-se que não há uma relação entre o menor tempo de processamento e os melhores desempenhos indicados anteriormente, visto que neste aspecto destacam-se os modelos *ZeroR* (0,1 segundo) e *OneR* (0.47 segundos), com os menores segundos de processamento, respectivamente, sendo que o *ZeroR* não fora citado em nenhum indicador de avaliação como um dos melhores desempenhos.

Assim como no experimento A, os algoritmos *J48* e o *Part* são modelos selecionados para o detalhamento dos respectivos resultados, visto que ambos destacaram-se em 4 (quatro) dos 7 (sete) indicadores de avaliação de desempenho, sendo os mesmos: classificação correta (*Correctly Classified Instances*), estatística Kappa (*Kappa statistic*), Erro Quadrado Médio (*Root mean squared error)* e Raiz do Erro Quadrado Relativo (*Root relative squared error).* Os resultados em detalhes para o referido experimento são indicados nas subseções a seguir:

4.4.2.1 J48

Com os seus parâmetros na forma padrão e tendo como atributo meta o *TipoDefic para o* presente experimento, o algoritmo J48 apresentou como resultado uma árvore de decisão com 11.150 folhas e tamanho 13.349. Apesar do seu extenso tamanho, é possível perceber pelo resultado gráfico gerado que a raiz da árvore é o atributo IndAprendiz, onde se desdobram os primeiros nós, representados por UF (para o IndAprendiz = sim) e o CNAEClasse (para o IndAprendiz = não). No terceiro nível identifica-se o atributo QtdHoraContrat (para IndAprendiz = não e CNAEClasse = C) e o atributo CBO2002Ocupacao para CNAEClasse = G (no mesmo ramo do IndAprendiz = não). Para as demais classes do CNAE no referido ramo não foi possível verificar os próximos nós ou folhas devido a sobreposição dos mesmos. Ainda no mesmo nível, identifica-se o atributo CBO2002Ocupacao para o UF = PR (ramo do IndAprendiz = sim), para as demais classes da UF não foi possível verificar os próximos nós ou folhas devido a sobreposição dos mesmos, assim como a continuação da mesma. No quarto nível do ramo do IndAprendiz = não, ainda é possível identificar o atributo UF para QtdHoraContrat= jornada 3, onde continua para o atributo UF e a partir da mesma, os Idade para o UF = RS e RaçaCor para UF = SC, sendo este o quinto nível. Para os demais ramos a partir do atributo CBO2002Ocupacao no terceiro nível (CNAEClasse = G e demais classes), não foi possível verificar os próximos nós ou folhas devido a sobreposição dos mesmos. Os atributos identificados estão conforme indica a Figura 28.

FIGURA 28 - IDENTIFICAÇÃO DA RAÍZ E DOS PRIMEIROS NÍVEIS/NÓS DA ÁRVORE DE DECISÃO DO J48 NO EXPERIMENTO B1 (PERFIL DAS PCD)



FONTE: A AUTORA COM BASE NOS RESULTADOS DA PESQUISA UTILIZANDO A FERRAMENTA *WEKA* (2017).

Os atributos identificados no topo da árvore de decisão resultante da aplicação do J48 no presente experimento que visa identificar o perfil dos PcD na região Sul do Brasil, são destaques por se constituírem os fatores que mais se correlacionam com o atributo meta selecionado (TipoDefic), ou seja, são os fatores importantes que levam ao perfil dos deficientes na região Sul do Brasil. A sobreposição dos nós e folhas geradas na continuação da árvore, resultado do seu tamanho extenso, impossibilitou uma análise completa, cuja árvore original pode ser consultada no Anexo D.

Visando explorar tais resultados sob uma outra perspectiva, a Figura 29 a seguir apresenta detalhadamente os valores pertinentes à acurácia de tal experimento:

FIGURA 29 - ACURÁCIA J48 - EXPERIMENTO B1 (PERFIL DAS PCD)

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0,253	0,091	0,453	0,253	0,325	0,203	0,672	0,397	AUDITIVA
0,850	0,616	0,519	0,850	0,645	0,258	0,683	0,615	FISICA
0,024	0,006	0,322	0,024	0,045	0,063	0,617	0,153	VISUAL
0,297	0,029	0,531	0,297	0,381	0,351	0,777	0,385	REABILITADO
0,029	0,001	0,417	0,029	0,055	0,108	0,658	0,048	MULTIPLA
0,454	0,041	0,591	0,454	0,513	0,464	0,819	0,504	Intelectual (Mental)
0,516	0,300	0,492	0,516	0,460	0,256	0,699	0,474	

FONTE: DADOS DA PESQUISA UTILIZANDO A FERRAMENTA *WEKA* (2017)

Conforme é possível verificar na Figura 29, os valores obtidos no *True Positive* (*TP Rate*) indicam uma probabilidade mais alta de um verdadeiro positivo na proporção de 85% na classe Física, seguido pela classe Intelectual (mental) (45%), o que nos possibilita identificar que que o padrão detectado acerta mais vezes nos casos de deficiência Física e Intelectual, respectivamente, do que com as deficiências Visual e Múltipla, que juntos apresentaram o menor TP representado por 0,2%.

Conforme já apresentado, os resultados obtidos em *False Positive* (*FP Rate*) indicam os dados classificados erroneamente como positivos, sendo a maior probabilidade encontrada deficiência Física com 61% e a menor em deficiência Múltipla - que possui 0%.

Como dito anteriormente, a precisão indica o valor da predição positiva (número de casos positivos por total de casos cobertos). Os resultados foram melhores para a classificação da deficiência Intelectual, apresentando 59% de precisão, enquanto a deficiência Visual apresentou 32%.

Visto que que corresponde ao valor da cobertura de casos, o resultado do *recall* (cobertura) é equivalente aos valores obtidos no verdadeiro positivo (TP).

Os valores obtidos no *F-Measure* indicam um desempenho de 64% para a classificação da deficiência Física e de 0,4% para a classificação da deficiência visual.

Os resultados obtidos em MCC demonstram um desempenho de 46% para a classificação da deficiência Intelectual. O *ROC Área* do classificador de deficiência Intelectual ficou em 81%, representando um bom resultado quando comparado com as demais classes, visto que valores próximos a 1 definem um classificador como ótimo. A pior classificação representada por 61% da deficiência Visual. Por fim, o *PRC Área* da deficiência Física ficou em 61%, enquanto a pior classificação foi da deficiência múltipla com 0,4%.

A matriz de confusão do referido algoritmo no presente experimento é indicada na Figura 30 a seguir:

FIGURA 30 - MATRIZ DE CONFUSÃO J48 - EXPERIMENTO B1 (PERFIL DAS PCD)

```
    a     b     c     d     e     f   <-- classified as
15036 38625   415  1973    24  3291 |     a = AUDITIVA
 9468 96298   595  3020    52  3900 |     b = FISICA
 3009 20707   643  1050    21  1211 |     c = VISUAL
 2504 14817   131  7553     9   414 |     d = REABILITADO
  349  2310    34   119   101   548 |     e = MULTIPLA
 2841 12717   179   500    35 13507 |     f = Intelectual (Mental)
```

FONTE: DADOS DA PESQUISA UTILIZANDO A FERRAMENTA *WEKA* (2017).

A matriz de confusão da Figura 30 possibilita um ângulo de visão diferente dos mesmos resultados, visto que os valores contidos na diagonal principal é a quantidade de registros classificada corretamente por classe, conforme explicitado anteriormente (vide Figura 20), é possível identificar que a classe do tipo de deficiência visual é a qual menos obteve seus registros classificados corretamente, visto que 98% dos seus registros encontram se fora do seu valor diagonal (valor 643, que por sua vez representa apenas 2% dos registros de deficiência visual). Por outro lado, possibilita observar que a classe do tipo de deficiência física é a qual obteve mais classificações corretas, o que representa 85% dos seus registros (97.591 na diagonal principal).

4.4.2.2 Part

Com os seus respectivos parâmetros no padrão (vide Figura 25), o algoritmo Part apresentou como resultado 12.457 regras tendo como atributo meta o *TipoDefic* - das quais destacam-se 11 na Tabela 8 a seguir, cujo critério de seleção fora o número de casos corretamente cobertos pela respectiva regra igual ou maior a 500 (indicado pela terceira e última coluna da tabela), devido ao volume de registros da base processada.

TABELA 8 - 11 REGRAS DESTACADAS DO ALGORITMO PART - EXPERIMENTO B1 (PERFIL DAS PCD)

(Continua)

REGRA - ALGORITMO PART	**INTERPRETAÇÃO**	**Casos cobertos corretamente/casos cobertos incorretamente**
CNAE20Classe = C AND UF = PR AND sexo = mas AND GrauInstrucao = Médio Completo AND QtdHoraContrat = jornada 3 AND RacaCor = BRANCA AND FaixaEmprInicioJan = DE 100 A 249: FISICA	*CNAE20Classe = Indústrias de transformação E UF = PR E sexo = masculino E GrauInstrucao = Médio Completo E QtdHoraContrat = acima de 29h33min semanais E RacaCor = branca E FaixaEmprInicioJan = de 100 a 249: FISICA*	896.63/496.04
UF = PR AND CNAE20Classe = C AND sexo = mas AND CBO2002Ocupacao = G AND Idade = adulto AND GrauInstrucao = Médio Completo AND TipoMovDesagregado = Desligamento por Demissão sem Justa Causa AND FaixaEmprInicioJan = 1000 OU MAIS AND TempoEmprego = periodo 1 AND RacaCor = BRANCA AND QtdHoraContrat = jornada 3: FISICA	*UF = PR E CNAE20Classe = Indústrias de transformação E sexo = masculino E CBO2002Ocupacao = Trabalhadores agropecuários, florestais, da caça e pesca E Idade = entre 25 e 59 anos E GrauInstrucao = Médio Completo E TipoMovDesagregado = Desligamento por Demissão sem Justa Causa E FaixaEmprInicioJan = 1000 ou mais E TempoEmprego = menor ou igual a 197 meses E RacaCor = branca E QtdHoraContrat = acima de 29h33min semanais: FISICA*	716.51/383.18
UF = RS AND Idade = adulto AND CNAE20Classe = C AND QtdHoraContrat = jornada 3 AND TipoMovDesagregado = Admissão por Reemprego AND FaixaEmprInicioJan = 1000 OU MAIS: FISICA	*UF = RS E Idade = entre 25 e 59 anos E CNAE20Classe = Indústrias de transformação E QtdHoraContrat = acima de 29h33min semanais E TipoMovDesagregado = Admissão por Reemprego E FaixaEmprInicioJan = 1000 ou mais: FISICA*	710.45/365.26
UF = RS AND Idade = adulto AND CNAE20Classe = C AND QtdHoraContrat = jornada 3 AND GrauInstrucao = Médio Completo AND RacaCor = BRANCA AND sexo = mas AND TipoMovDesagregado = Desligamento por Demissão sem Justa Causa: FISICA	*UF = RS E Idade = entre 25 e 59 anos E CNAE20Classe = Indústrias de transformação E QtdHoraContrat = acima de 29h33min semanais E GrauInstrucao = Médio Completo E RacaCor = branca E sexo = masculino E TipoMovDesagregado = Desligamento por Demissão sem Justa Causa: FISICA*	693.61/363.51
UF = PR AND CNAE20Classe = C AND sexo = mas AND Idade = adulto AND GrauInstrucao = Médio Completo AND RacaCor = BRANCA AND TipoMovDesagregado = Admissão por Reemprego AND FaixaEmprInicioJan = 1000 OU MAIS AND QtdHoraContrat = jornada 3: FISICA	*UF = PR E CNAE20Classe = Indústrias de transformação E sexo = masculino E Idade = entre 25 e 59 anos E GrauInstrucao = Médio Completo E RacaCor = branca E TipoMovDesagregado = Admissão por Reemprego E FaixaEmprInicioJan = 1000 ou mais E QtdHoraContrat = acima de 29h33min semanais: FISICA*	691.03/415.55

(Continuação)

REGRA - ALGORITMO PART	INTERPRETAÇÃO	Casos cobertos corretamente/casos cobertos incorretamente
UF = SC AND CNAE20Classe = A AND GrauInstrucao = Fundamental Completo AND RacaCor = BRANCA AND CBO2002Ocupacao = F AND sexo = mas: REABILITADO	*UF = SC E CNAE20Classe = a E GrauInstrucao = Fundamental Completo E RacaCor = branca E CBO2002Ocupacao = F E sexo = masculino: REABILITADO*	605.0/32.0
CNAE20Classe = C AND UF = PR AND sexo = mas AND GrauInstrucao = Médio Completo AND RacaCor = BRANCA AND CBO2002Ocupacao = G AND TipoMovDesagregado = Desligamento a Pedido: AUDITIVA	*CNAE20Classe = Indústrias de transformação E UF = PR E sexo = masculino E GrauInstrucao = Médio Completo E RacaCor = branca E CBO2002Ocupacao = Trabalhadores agropecuários, florestais, da caça e pesca E TipoMovDesagregado = Desligamento a Pedido: AUDITIVA*	604.87/370.55
UF = RS AND Idade = adulto AND CNAE20Classe = C AND QtdHoraContrat = jornada 3 AND TipoMovDesagregado = Admissão por Reemprego AND GrauInstrucao = 6ª a 9ª Fundamental AND CBO2002Ocupacao = G AND FaixaEmprInicioJan = DE 500 A 999 AND RacaCor = BRANCA: FISICA	*UF = RS E Idade = entre 25 e 59 anos E CNAE20Classe = Indústrias de transformação E QtdHoraContrat = acima de 29h33min semanais E TipoMovDesagregado = Admissão por Reemprego E GrauInstrucao = 6ª a 9ª Fundamental E CBO2002Ocupacao = Trabalhadores agropecuários, florestais, da caça e pesca E FaixaEmprInicioJan = de 500 a 999 E RacaCor = branca: FISICA*	546.78/274.35
UF = RS AND CNAE20Classe = G AND GrauInstrucao = Médio Completo AND QtdHoraContrat = jornada 3 AND sexo = fem AND CBO2002Ocupacao = D AND RacaCor = BRANCA AND TipoMovDesagregado = Admissão por Reemprego: FISICA	*UF = RS E CNAE20Classe = Comércio; reparação de veículos automotores e motocicletas E GrauInstrucao = Médio Completo E QtdHoraContrat = acima de 29h33min semanais E sexo = feminino E CBO2002Ocupacao = Técnicos de nível médio E RacaCor = branca E TipoMovDesagregado = Admissão por Reemprego: FISICA*	539.49/242.54
UF = PR AND CNAE20Classe = C AND sexo = fem AND RacaCor = BRANCA AND TipoMovDesagregado = Admissão por Reemprego AND GrauInstrucao = Médio Completo AND FaixaEmprInicioJan = 1000 OU MAIS AND QtdHoraContrat = jornada 3: AUDITIVA	*UF = PR E CNAE20Classe = Indústrias de transformação E sexo = feminino E RacaCor = branca E TipoMovDesagregado = Admissão por Reemprego E GrauInstrucao = Médio Completo E FaixaEmprInicioJan = 1000 ou mais E QtdHoraContrat = acima de 29h33min semanais: AUDITIVA*	520.59/260.6

(Continuação)

REGRA - ALGORITMO PART	INTERPRETAÇÃO	Casos cobertos corretamente/casos cobertos incorretamente
CNAE20Classe = C AND Idade = adulto AND UF = SC AND RacaCor = BRANCA AND CBO2002Ocupacao = G AND FaixaEmprInicioJan = 1000 OU MAIS AND TipoMovDesagregado = Admissão por Reemprego: REABILITADO	*CNAE20Classe = Indústrias de transformação E Idade = entre 25 e 59 anos E UF = SC E RacaCor = branca E CBO2002Ocupacao = Trabalhadores agropecuários, florestais, da caça e pesca E FaixaEmprInicioJan = 1000 ou mais E TipoMovDesagregado = Admissão por Reemprego: REABILITADO*	515.77/284.83

FONTE: DADOS DA PESQUISA UTILIZANDO A FERRAMENTA *WEKA* (2017).

Devido a quantidade de atributos selecionados para tal experimento, é evidente que as regras apresentadas são bem específicas, conforme indica a coluna Interpretação. Com cobertura de mais de 900 regras corretamente classificadas, identifica que muitos deficientes físicos que moram no PR possuem ensino médio completo, são masculinos e se declaram brancos, trabalham acima de 30h semanais em empresas cuja faixa de funcionários é de 100 a 249, sendo esta a primeira regra destacada, onde percebe-se um perfil bem específico delineado.

Visando explorar tais resultados com outra perspectiva, a Figura 31 a seguir apresenta detalhadamente os valores pertinentes à acurácia de tal experimento:

FIGURA 31 - ACURÁCIA PART - EXPERIMENTO B1 (PERFIL DAS PCD)

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0,304	0,127	0,416	0,304	0,351	0,199	0,675	0,392	AUDITIVA
0,785	0,546	0,530	0,785	0,632	0,248	0,684	0,625	FISICA
0,046	0,018	0,230	0,046	0,077	0,061	0,615	0,154	VISUAL
0,333	0,037	0,494	0,333	0,398	0,354	0,795	0,405	REABILITADO
0,034	0,001	0,301	0,034	0,062	0,098	0,637	0,052	MULTIPLA
0,453	0,046	0,563	0,453	0,502	0,448	0,820	0,515	Intelectual (Mental)
0,505	0,280	0,470	0,505	0,464	0,249	0,701	0,480	

FONTE: DADOS DA PESQUISA UTILIZANDO A FERRAMENTA *WEKA* (2017).

Conforme é possível verificar na Figura 31, os valores obtidos no *True Positive* (*TP Rate*) indicam uma probabilidade mais alta de um verdadeiro positivo na proporção de 78% na classe Física, seguido pela classe Intelectual (mental) (45%), o que nos possibilita identificar que que o padrão detectado acerta mais vezes nos casos

de deficiência Física e Intelectual, respectivamente, do que com as deficiências Visual e Múltipla, que juntos apresentaram o menor TP representado por 0,2% e 0,3%, respectivamente.

Conforme já apresentado, os resultados obtidos em *False Positive* (*FP Rate*) indicam os dados classificados erroneamente como positivos, sendo a maior probabilidade encontrada deficiência Física com 54% e a menor em deficiência Múltipla - que possui 0,0%.

Como dito anteriormente, a precisão indica o valor da predição positiva (número de casos positivos por total de casos cobertos). Os resultados foram melhores para a classificação da deficiência Intelectual, apresentando 56% de precisão, enquanto a deficiência Visual apresentou 23%.

Visto que que corresponde ao valor da cobertura de casos, o resultado do *recall* (cobertura) é equivalente aos valores obtidos no verdadeiro positivo (TP).

Os valores obtidos no *F-Measure* indicam um desempenho de 63% para a classificação da deficiência Física e de 0,6% para a classificação da deficiência múltipla.

Os resultados obtidos em MCC demonstram um desempenho de 44% para a classificação da deficiência Intelectual. O *ROC Área* do classificador de deficiência Intelectual ficou em 82%, representando um bom resultado quando comparado com as demais classes, visto que valores próximos a 1 definem um classificador como ótimo. A pior classificação representada por 61% da deficiência Visual. Por fim, o *PRC Área* da deficiência intelectual ficou em 51%, enquanto a pior classificação foi da deficiência múltipla com 0,5%.

A matriz de confusão do referido algoritmo no presente experimento é indicada na Figura 32 a seguir:

FIGURA 32 - MATRIZ DE CONFUSÃO PART - EXPERIMENTO B1 (PERFIL DAS PCD)

```
    a     b     c     d     e     f   <-- classified as
18023 34322  1058  2459    57  3445 |     a = AUDITIVA
13764 88910  2014  4020    94  4531 |     b = FISICA
 4133 18411  1228  1391    50  1428 |     c = VISUAL
 3159 12858   411  8469    23   508 |     d = REABILITADO
  539  2049    76   129   119   549 |     e = MULTIPLA
 3672 11350   545   665    53 13494 |     f = Intelectual (Mental)
```

FONTE: DADOS DA PESQUISA UTILIZANDO A FERRAMENTA *WEKA* (2017).

A matriz de confusão da Figura 32 possibilita um ângulo de visualização diferente para os mesmos resultados, visto que os valores contidos na diagonal principal é a quantidade de registros classificada corretamente por classe, conforme explicitado anteriormente (vide Figura 20), é possível identificar que a classe do tipo de deficiência visual é a qual menos obteve seus registros classificados corretamente, visto que 95% dos seus registros encontram se fora do seu valor diagonal (valor 1228, que por sua vez representa apenas 5% dos registros de deficiência visual). Por outro lado, possibilita observar que a classe do tipo de deficiência física é a qual obteve mais classificações corretas, o que representa 78% dos seus registros (88.910 na diagonal principal).

4.6 MINERAÇÃO DE DADOS – EXPERIMENTO 2

Dado que o processo é interativo (vide Figura 9), onde o usuário pode intervir e controlar a sequência das etapas, retorna-se novamente à etapa de mineração de dados com o objetivo de obter resultados genéricos e mais satisfatórios do que os experimentos anteriores, de forma a melhor responder as questões de pesquisa e consequentemente apontar uma aplicação destes no cotidiano. Assim como apresentado anteriormente, os detalhes dos resultados atingidos se dará por experimento nas próximas subseções, que por sua vez, corresponde ao objetivo requerido:

4.6.1 Experimento A2 - Padrão no mercado formal das PcD

Conforme já dito anteriormente, visando obter resultados melhores e genéricos, de forma a apontar a aplicação destes no cotidiano, na referida etapa selecionou-se somente os atributos principais, visto que são os que mais impactam no referido objetivo que se deseja descobrir com a referida mineração (algum padrão no mercado formal das PcD): CBO2002Ocupacao, CNAE20Classe, FaixaEmprInicioJan, GrauInstrucao, QtdHoraContrat, Idade, SalarioMensal, TipoDefic e TipoMovDesagregado.

Descartou-se o atributo UF de modo a obter resultados mais homogêneos e não desagregado por estado. Os algoritmos aplicados foram os mesmos da etapa anterior de mineração a fim de manter um patamar de comparação dos resultados, que podem ser consultados na Tabela 9 a seguir:

TABELA 9 - INDICADORES DE DESEMPENHO DOS RESULTADOS DO EXPERIMENTO A2 (PADRÃO NO MERCADO FORMAL DAS PCD)

		Correctly Classified Instances (%)	Kappa statistic	Mean absolute error	Root mean squared error	Relative absolute error (%)	Root relative squared error (%)	Tempo de processa-mento
RULES	*DecisionTable*	42.25	0.0727	0.1309	0.2549	97.72	98.48	48.16
	JRip	40.84	0.0014	0.1338	0.2587	99.93	99.97	19.41
	OneR	41.04	0.0098	0.1072	0.3274	80.04	126.52	0.14
	Part	42.44	0.0861	0.1278	0.2546	95.45	98.39	404.7
	ZeroR	40.80	0	0.1339	0.2588	100	100	0.04
TREES	DecisionStump	40.80	0	0.1328	0.2577	99.17	99.58	0.17
	HoeffdingTree	42.03	0.0769	0.128	0.2551	95.56	98.57	4.46
	J48	42.70	0.0773	0.1292	0.2549	96.44	98.50	3.04
	RandomTree	40.26	0.0718	0.1281	0.2644	95.62	102.18	1.42
	REPTree	42.31	0.0818	0.1281	0.2555	95.67	98.71	3.95

FONTE: DADOS DA PESQUISA UTILIZANDO A FERRAMENTA *WEKA* (2017).

Assim como no primeiro experimento com mesmo objetivo (A1), o algoritmo *PART* repete-se como o algoritmo de melhor desempenho por atingir tal resultado em cinco dos sete indicadores de avaliação: *correctly classified instances, Kappa Statistica, Mean absolute error, relative absolute error* e *root relative squared error*. O *J48* também consolida-se como um dos melhores novamente, mas empata-se com o *DecisionTable* e o *OneR,* visto que os três atingiram o resultado em dois dos cinco indicadores. Porém, o desempenho de tais algoritmos não se alavancaram quando comparados com os resultados do primeiro experimento com o mesmo objetivo (subseção 4.5.1), ficando inferior aos resultados dos mesmos em todos os indicadores de avaliação (vide Tabela 5), ou seja, mesmo com a seleção de poucos, que porventura também são considerados os principais, mostra que para a presente base de dados, os resultados não dependem do número de atributos selecionados e nem se potencializa com a seleção de poucos destes. Diante disso, descarta-se a necessidade de entrar detalhes dos algoritmos com os melhores desempenhos de tais

resultados, prevalecendo os resultados apresentados dos algoritmos destacados como os melhores em desempenho do Experimento A1 (vide seção 4.5.1).

4.6.1 Experimento B2 - Perfil das PcD

Conforme já dito anteriormente, visando obter resultados melhores e genéricos, de forma a apontar a aplicação destes no cotidiano, na referida etapa selecionou-se somente os atributos principais, visto que são os que mais impactam no referido objetivo que se deseja descobrir com a referida mineração (perfil das PcD da região Sul): GrauInstrucao, Idade, RacaCor, sexo, TipoDefic e TipoMovDesagregado.

Descartou-se o atributo UF de modo a obter resultados mais homogêneos e não desagregado por estado. Os algoritmos aplicados foram os mesmos da etapa anterior de mineração a fim de manter um patamar de comparação dos resultados, que podem ser consultados na Tabela 10 a seguir:

TABELA 10 - INDICADORES DOS RESULTADOS EXPERIMENTO B2 (PERFIL DAS PCD)

		Correctly Classified Instances (%)	Kappa statistic	Mean absolute error	Root mean squared error	Relative absolute error (%)	Root relative squared error (%)	Tempo de processa-mento
RULES	*DecisionTable*	45.87	0.0774	0.2317	0.3403	96.50	98.22	5.09
	JRip	44.05	0.0043	0.2397	0.3462	99.86	99.93	5.78
	OneR	44.24	0.0189	0.1859	0.4311	77.41	124.43	0.08
	Part	45.90	0.0779	0.2318	0.3405	96.53	98.27	5.96
	ZeroR	43.92	0	0.2401	0.3465	100	100	0.05
TREES	*DecisionStump*	43.92	0	0.2382	0.3451	99.20	99.6	0.1
	HoeffdingTree	45.84	0.0776	0.2317	0.3406	96.51	98.29	1.53
	J48	45.90	0.0766	0.232	0.3406	96.63	98.30	0.58
	RandomTree	45.89	0.0787	0.2311	0.3401	96.26	98.17	0.41
	REPTree	45.87	0.0775	0.2319	0.3406	96.61	98.31	1.38

FONTE: DADOS DA PESQUISA UTILIZANDO A FERRAMENTA *WEKA* (2017).

Diferente dos experimentos apresentados anteriormente, desta vez o algoritmo *RandomTree* consolida-se como o algoritmo de melhor desempenho, por atingir tal feito em cinco dos sete indicadores de avaliação de desempenho: *Kappa*

Statistica, Mean absolute error, root mean squared error, relative absolute error e *root relative squared erro.* Na sequência se destacam os algoritmos *DecisionTable* e *Part* por atingirem melhores desempenho em dois dos sete indicadores de cada um. Porém, assim como no Experimento A2, é possível verificar que o desempenho de tais algoritmos não se potencializaram quando comparados com os resultados obtidos no experimento B1 que possui o mesmo objetivo (vide Tabela 7), ficando inferior aos mesmos, em todos os indicadores de avaliação, mesmo com a seleção de poucos atributos, que porventura são considerados os principais ao presente objetivo, mostrando que para a presente base de dados, os resultados não dependem do número de atributos selecionados e nem se alavanca com a seleção de poucos destes. Diante disso, descarta-se a necessidade de entrar detalhes dos algoritmos com os melhores desempenhos de tais resultados, prevalecendo os resultados apresentados dos algoritmos destacados como os melhores em desempenho do Experimento B1 (vide seção 4.5.2).

4.7 CONHECIMENTO DESCOBERTO

Visto que os resultados obtidos com a primeira etapa de mineração de dados (A1 e B1) não foram satisfatórios estatisticamente, além de não ser possível uma análise completa dos mesmos – pelo tamanho da árvore de decisão e o grande volume de regras geradas, sendo as mesmas bem específicas, o que impossibilitou o apontamento de aplicação das mesmas no cotidiano, realizou-se uma nova etapa de mineração de dados, onde selecionou-se poucos atributos para a referida mineração, sendo estes os considerados principais para cada objetivo proposto e descartando o atributo UF, de modo a obter resultados mais homogêneos e não desagregado por estado. Porém, os novos resultados não atingiram a expectativa e também não foram satisfatórios, ou seja, estatisticamente os novos resultados não foram superior aos resultados obtidos com a primeira etapa de mineração realizada, mostrando que para a presente base de dados, selecionar apenas poucos atributos ou os principais, não potencializa os resultados. Diante disso, os resultados obtidos com a primeira etapa se consolidam como resultados oficiais, sendo o foco para a análise a ser realizada ao longo do presente capítulo.

É possível notar que na referida etapa, o experimento B apresentou melhores resultados sobre a acurácia dos dados sobre o experimento A, de maneira geral. Apesar dos seus respectivos resultados serem bem mais extenso, seja o tamanho da árvore de decisão ou o total de regras geradas, o que dificulta a validação e análise dos mesmos, tal fato é perceptível na Tabela 11, onde compara-se o desempenho dos algoritmos em ambos os experimentos e destaca-se o melhor desempenho do experimento B em 5 dos 7 indicadores de avaliação de desempenho dos resultados:

TABELA 11 - COMPARAÇÃO DE DESEMPENHO - EXPERIMENTOS A1 X B1

	Experimento A		Experimento B	
	J48	**PART**	**J48**	**PART**
Correctly Classified Instances (%)	44.10	41.99	51.60	50.48
Kappa statistic	0.1055	0.1062	0.2416	0.2453
Mean absolute error	0.1265	0.1252	0.208	0.2034
Root mean squared error	0.2531	0.2565	0.3264	0.3286
Relative absolute error (%)	94.48	93.49	86.64	84.73
Root relative squared error (%)	97.81	99.11	94.20	94.83
Tempo de processa-mento	256.98	2977.14	9.17	2209.39

FONTE: DADOS DA PESQUISA UTILIZANDO A FERRAMENTA *WEKA* (2017).

Os algoritmos foram selecionados com base nos seus respectivos desempenhos e o fato de coincidentemente serem os mesmos para ambos os experimentos, motivou tal análise: O experimento B, cujo objetivo era identificar o perfil das PcD na região do Sul, teve a melhor taxa de dados classificados corretamente (*correctly classified instances*), em ambos os algoritmos acima de 50%, sendo as mesmas bem superiores ao do Experimento A, cujo objetivo era identificar padrões de desligamentos/admissão para os PcD da região Sul. Apesar dos resultados inconclusivos, para o experimento B identifica-se nível de concordância e/ou ligação entre os dados superior ao obtido com o experimento A. Os próximos quatro indicadores são pertinentes a predição do modelo e observa-se um empate, visto que o experimento A se destaca nos dois primeiros, possuindo a média do erro da predição

em ambos os algoritmos quando comparados com o experimento B (*mean absolute error)* e consequentemente tal resultado reflete no próximo indicador (r*oot mean squared error)* por ser uma derivada do *Mean absolute error*, ´pois é a média da raiz quadrada da diferença entre o valor predito e o valor real. Já nos dois últimos indicadores relacionados a predição, o experimento B se consolida novamente como o de melhor desempenho, indicando uma maior precisão do modelo sobre o experimento A no indicador *relative absolute error*, visto que os seus valores são inferiores em ambos os algoritmos quando comparados e os valores mais baixos significam uma maior precisão do referido modelo. O mesmo reflete no indicador *Root relative squared error*, cujos valores também são inferiores em ambos os algoritmos, indicando que a predição fora superior do que os dados reais.

Por meio da árvore de decisão gerada graficamente em ambos os experimentos, foi possível identificar a raiz e alguns nós/atributos na sequência, sendo estes o topo nas referidas árvore, indicando que são os fatores que mais se correlacionam com o objetivo desejado no referido experimento:

- EXPERIMENTO A: TempoEmprego em primeiro nível, seguido do IndAprendiz e CNAEClasse no segundo nível, Idade em terceiro nível, HoraContrat e CNAEClasse em quarto nível são fatores determinantes que levam ao padrão de desligamento/admissão na referida região de análise (vide Figura 22). A sobreposição das nós/folhas em muitos ramos impossibilitaram uma análise completa e identificação dos próximos atributos. A árvore de decisão completa pode ser consultada no ANEXO C.
- EXPERIMENTO B: IndAprendiz em primeiro nível, UF e CNAEClasse em segundo nível, seguido dos atributos QtdHorasContrat e CBO2002Ocupação em terceiro nível, UF em quarto nível, Idade e RaçaCor em quinto nível são fatores determinantes que determinam o perfil do PcD na referida região de análise (vide Figura 28). A sobreposição das nós/folhas em muitos ramos impossibilitaram uma análise completa e identificação dos próximos atributos. A árvore de decisão completa pode ser consultada no ANEXO D.

Já especificamente sobre o algoritmo PART, não foi possível um apontamento de aplicação no cotidiano para as regras geradas em ambos os experimentos (vide Tabelas 6 e 8), visto que as mesmas foram bem específicas devido a seleção de todos os atributos.

Conforme já dito, os algoritmos foram selecionados com base nos seus desempenhos com relação aos demais, sendo evidente que os algoritmos explorados em ambos os experimentos - J48 e PART, não apontam soluções completas e/ou válidas para a presente base de dados, onde identifica a necessidade de aplicar algoritmos mais complexos, fugindo do escopo delimitado da presente pesquisa (algoritmos do tipo *rules* e *tree*) e/ou relacionar com outra base de dados e/ou inserção de mais atributos, visando atingir resultados mais satisfatórios para análise e/ou informações úteis.

Castro e Ferrari (2016, p. 259) afirmam que “[...] não existe um classificador que seja melhor que todos os outros para todos os problemas de classificação”, ou seja, cada algoritmo possui suas singularidades e respectiva importância, de modo que pode-se afirmar que não existe o melhor método e sim o mais satisfatório para cada base de dados, pois o desempenho de um mesmo algoritmo sofre alterações de acordo com a base de dados analisada. Ora, não são satisfatórios para a presente base de estudo, mas para uma outra base de dados distinta e om outros objetivos, o resultado pode ser diferente, podendo até serem os melhores indicados pra a classificação.

Também foi possível obter algum conhecimento válido apenas com base na estatística descritiva, relacionados a seguir:

- Os deficientes físicos foram os mais responsáveis pela movimentação dos PcD na região Sul do Brasil dentro do período analisado, seguido dos deficientes auditivos (vide Gráfico 11). Tal informação é uma reflexão da distribuição dos tipos de deficiência na referida região (vide Gráfico 10);
- Na região Sul do Brasil, Admissão por reemprego é o motivo mais predominante para os registros de admissão (com representação de 83,30%), enquanto Desligamento por Demissão sem Justa Causa e Desligamento a Pedido são os motivos mais predominantes quando se

trata de desligamento (com 43,53% e 37,42% de representação, respectivamente). Tal padrão reflete para o tipo de sexo também, em ambas as situações (vide Gráficos 21 e 22);

- Quando se trata do saldo de movimentação, identificou-se que a região Sul seguiu a mesma tendência do Brasil dentro do período analisado (vide Gráfico 19), evidenciando que se o cenário econômico brasileiro está favorável, consequentemente isso refletirá no Sul com a geração de mais empregos e vice-versa;
- Já quando se trata do saldo de deficientes, evidenciou que a região Sul possuiu um saldo médio superior ao restante do Brasil (vide Gráfico 20) dentro do período analisado, exceto no ano de 2015.

5 CONSIDERAÇÕES FINAIS

Big Data, *Open Data* e democracia são assuntos que estão intimamente relacionados, embora tal relação não seja tão clara para a maioria das pessoas. Sabe-se que é papel dos órgãos públicos oferecer transparência em suas gestões e o acesso não só às informações aos cidadãos, mas também ao contexto variado e sempre atualizado de dados (*Big Data*) públicos, que é o papel que o Portal Brasileiro de Dados Abertos (PBDA) vem realizando.

Dentro do contexto do *Big Data*, os dados só tem valor quando analisados e

> [...] como os computadores permitiram que os humanos reunissem mais dados do que podemos digerir, é natural recorrer a técnicas computacionais para nos ajudar desenterrar padrões e estruturas significativas dos enormes volumes de dados. “(FAYYAD et al, 1996, p. 38)

O governo em sua plenitude, juntamente com as instituições privadas e acadêmicas, é um dos responsável pela vasta produção de dados, que quando disseminados, proporcionam ao “cidadão um melhor entendimento do governo, no acesso aos serviços públicos, no controle das contas públicas e na participação no planejamento e desenvolvimento das políticas públicas.” (Brasil, Portal Brasileiro de Dados Abertos). Por outro lado, “as empresas usam dados para ganhar vantagem competitiva, aumentar a eficiência e fornecer serviços mais valiosos aos clientes. (FAYYAD et al, 1996, p.38).

Diante de tal necessidade, a mineração de dados é uma ferramenta determinante, que muito tem a contribuir com a análise de dados abertos, seja para o bem ao cidadão comum, cujo resultado em muito pode contribuir com as ações já citadas anteriormente ou seja para conhecer o seu público alvo, descobrir um nicho de mercado e/ou construir índices baseados em dados sociais para as empresas privadas.

5.1 VALIDAÇÃO DOS OBJETIVOS ESTABELECIDOS

A fim de atingir o objetivo geral do presente trabalho, onde determinou-se a buscar padrões do mercado formal de trabalho das PcD na região Sul e identificar o

perfil das mesmas através de aplicação das técnicas de árvore de decisão e regras da tarefa de classificação dentro da etapa da mineração de dados do KDD, projetou-se quatro objetivos específicos:

O primeiro objetivo específico determinado, que consiste em aplicar técnicas de mineração, do tipo classificação, na base de dados de deficiente da região Sul do Brasil, provenientes do CAGED foi alcançado por meio das etapas práticas de coleta dos dados da referida base, seleção dos dados pertinentes ao referido objetivo (dados de PcD e da região Sul do Brasil), a transformação dos seus valores para melhor aceitação na ferramenta *WEKA* (visto que originalmente os valores são numéricos, característica esta que não ativa a maioria e os mais relevantes algoritmos de classificação na referida ferramenta), permitindo assim, a efetiva mineração de dados em duas etapas, onde aplicou-se os algoritmos de regras e árvore de decisão.

O segundo objetivo que consiste em identificar os melhores algoritmos para a referida base de dados, com base nos seus desempenhos foi alcançado com sucesso em razão do objetivo anterior ter sido alcançado. Visto que foi possível aplicar os algoritmos das técnicas determinadas e tendo dois objetivos distintos pra a etapa de mineração de dados – o que consequentemente motivou dois experimentos independentes, foi possível avaliar o desempenho dos mesmos por experimento, onde chegou-se na consideração de que dentro do escopo determinado (tarefa de classificação), os algoritmos J48 e PART, dos tipos árvore de decisão e regras, respectivamente, são os mais adequados para a aplicação na referida base de dados quando comparado com os demais algoritmos em termos de desempenho, apesar d que estatisticamente tais resultados são satisfatórios.

O terceiro objetivo determinado, que consiste em descrever o mercado de trabalho formal das pessoas com deficiências na região Sul por meio da aplicação de algoritmos de classificação sobre a referida base de dados foi atingido parcialmente: estatisticamente os resultados não foram satisfatórios, mas a visualização gráfica da árvore de decisão gera por meio da aplicação do J48 em ambos os experimentos permitiu identificar alguns atributos, que são os mais relacionados e determinam os resultados do objetivo da referida aplicação. Porém devido a sobreposição dos nós/folhas nas árvores, resultado do seu extenso tamanho, não foi possível realizar uma análise completa. As regras resultadas da aplicação do PART foram bem específicas, o que não permitiu um apontamento prático no cotidiano e assim, descrever o mercado de trabalho dos PcD. Parcialmente porque tal objetivo não fora

alcançado por completo por meio da mineração de dados em si, mas por intermédio da estatística descritiva foi possível identificar informações válidas e assim, descrever o mercado de trabalho – vide subseção 4.7 para maiores detalhes.

O quarto e último objetivo, que consiste em analisar e apresentar os padrões obtidos da mineração, apontando aplicações destes no cotidiano, intrinsecamente relacionado ao objetivo anterior, que conforme já esclarecido, não foi possível apontar a aplicação dos resultados obtidos no cotidiano, considerando que o presente objetivo não foi alcançado em razão dos resultados da mineração de dados.

5.2 CONTRIBUIÇÕES

Mesmo com limitações impostas, tais como restrição de tempo e hardware, o desenvolvimento da presente pesquisa permitiu identificar a relevância em correlacionar dados abertos e mineração de dados e os seus benefícios para o cidadão. Se os resultados obtidos não fossem tão complexos para análise e não requeresse testes e validações com outros algoritmos fora os determinado pelo presente escopo, haveria possibilidade de contribuir com as políticas públicas atuais e assim, trabalhar em função de uma melhor e maior inclusão das PcD no mercado de trabalho.

Como o governo disponibiliza uma gama incontável de bases de dados, sobre todos os demais setores da sociedade, outros objetivos podem ser alcançados com outros tipos de bases públicas e assim, atingir a expectativa, visto que o conhecimento descoberto será de valor social, contribuindo com as políticas públicas brasileiras, dado o valor que os dados coletados e armazenados pelo governo possuem.

Espera-se que o presente trabalho contribua para uma reflexão acerca da forma com que tais bases de dados públicas vem sendo utilizadas e o como o seu uso efetivo com ferramentas e aplicações corretas em muito pode beneficiar a sociedade.

5.3 SUGESTÕES DE TRABALHOS FUTUROS

Para trabalhos futuros relacionados ao tema da presente pesquisa, recomenda-se as seguintes sugestões:

a) Mineração dos dados totais do CAGED, de forma a buscar padrões ou tendências do mercado formal de trabalho em um contexto geral, sem segregação por estado e/ou região;
b) Aplicação das técnicas de mineração de dados em outras sub bases, podendo a mesma ser segregada por outras regiões do Brasil não exploradas ou estado, de forma a buscar padrões e/ou tendências do mercado formal de trabalho dentro de um contexto ainda mais específico;
c) Para a referida base de estudo, onde selecionou-se os dados pertinentes ao Sul ou para as demais bases sugeridas anteriormente, explorar outras tarefas e técnicas de mineração de dados além do que fora explorado, de forma a buscar padrões válidos ou novos conhecimentos.
d) Por fim, explorar outros tipos de dados abertos para aplicação das técnicas de mineração de dados, sejam dados de outros setores da sociedade ou não, de forma buscar novos padrões ou algum conhecimento de valor social em benefício da nossa sociedade.

REFERÊNCIAS

AMORIM, T. **Conceitos, técnicas, ferramentas e aplicações de Mineração de Dados para gerar conhecimento a partir de bases de dado.** 2006. 50 f. TCC (Graduação) - Curso de Ciência da Computação, Universidade Federal de Pernambuco, Recife, 2006.

BERRY, M. J. A.; LINOFF, G. S. **Data Mining Techniques:** for marketing, sales, and customer Relationship Management. 2. ed. Indianapolis: Wiley Publishing, 2004. 643 p.

BRASIL. Constituição (1988). **Constituição da República Federativa do Brasil de 1988**. Brasília, DF, 1988. Disponível em: <http://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm>. Acesso em: 03 abr. 2017.

BRASIL. **Convenção sobre os Direitos das Pessoas com Deficiência**: Protocolo Facultativo à Convenção sobre os Direitos das Pessoas com Deficiência: Decreto Legislativo nº 186, de 09 de julho de 2008: Decreto nº 6.949, de 25 de agosto de 2009. 4ª Ed., rev. e atual. Brasília: Secretaria de Direitos Humanos, 2010. 100p. Disponível em: <http://www.pessoacomdeficiencia.gov.br/app/sites/default/files/publicacoes/convencaopessoascomdeficiencia.pdf>. Acesso em 18 abr. 2017.

BRASIL. **Decreto Nº 3.298, de 20 de Dezembro de 1999.** Brasília, DF. Disponível em: <http://www.planalto.gov.br/ccivil_03/decreto/d3298.htm>. Acesso em: 03 abr. 2017.

BRASIL. **Decreto Nº 6.949, de 25 de Agosto de 2009.** Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2007-2010/2009/decreto/d6949.htm>. Acesso em: 03 abr. 2017.

BRASIL. **Decreto Nº 7.724, de 16 de Maio de 2012.** Disponível em:< http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2012/decreto/d7724.htm>. Acesso em 03 abr. 2017

BRASIL. **IBGE:** Instituto Brasileiro de Geografia e Estatística**.** Disponível em: < https://www.ibge.gov.br/estatisticas-novoportal/sociais/populacao.html >. Acesso em: 10 ago. 2011.

BRASIL. **Lei Nº 7.853, de 24 de Outubro de 1989**. Disponível em: <http://www.planalto.gov.br/ccivil_03/leis/L7853.htm>. Acesso em: 03 abr. 2017.

BRASIL. **Lei Nº 12.527, de 18 de Novembro de 2011.** Brasília, DF. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm>. Acesso em: 29 mar. 2012.

BRASIL. **Parceria para governo aberto**. Disponível em: <http://www.governoaberto.cgu.gov.br/no-brasil/planos-de-acao-1/1o-plano-de-acao-do-brasil>. Acesso em: 01 maio 2017.

BRASIL. **Portal Brasileiro de Dados Abertos**. Disponível em: <http://dados.gov.br/paginas/dados-abertos>. Acesso em: 06 mar. 2017.

BRASIL. Ministério do Trabalho. **Microdados RAIS e CAGED.** 2016. Disponível em: <http://pdet.mte.gov.br/microdados-rais-e-caged>. Acesso em: 06 mar. 2017.

BRASIL. Ministério do Trabalho. **O que é CAGED?** 2016. Disponível em: <http://pdet.mte.gov.br/o-que-e-caged>. Acesso em: 06 mar. 2017.

BRASIL. Secretaria Nacional de Promoção dos Direitos da Pessoa com Deficiência. **Cartilha Do Censo 2010:** Pessoas com Deficiência. 2012. Disponível em: <http://www.pessoacomdeficiencia.gov.br/app/sites/default/files/publicacoes/cartilha-censo-2010-pessoas-com-deficienciareduzido.pdf>. Acesso em: 03 abr. 2017.

CAMILO, C. O.; SILVA, J. C. da. **Mineração de dados**: conceitos, tarefas, métodos e ferramentas. Goiânia: Instituto de Informática, 2009. Disponível em: <http://www.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_001-09.pdf>. Acesso em: 09. mai. 2017.

CARVALHO, L. A. V. de. **Data Mining**: a mineração de dados no marketing, medicina, economia, engenharia e administração. São Paulo: Editora Érica, 2001.

CASTRO, L. N; FERRARI, D. G. **Introdução à mineração de dados:** conceitos básicos, algoritmos e aplicações. São Paulo: Saraiva, 2016.

DADOS ABERTOS GOVERNAMENTAIS. **Manual dos dados abertos:** governo. Disponível em: < http://www.w3c.br/pub/Materiais/PublicacoesW3C/Manual_Dados_Abertos_WEB.pdf > Acesso em 02. Out. 2017.

DAVENPORT, T. H. **Ecologia da informação:** por que só a tecnologia não basta para o sucesso na era da informação. São Paulo: Futura, 1998. 312 p.

ESTERMANN, B. Diffusion of open data and crowdsourcing among heritage institutions: results of a pilot survey in switzerland. **Journal of Theoretical and Applied Electronic Commerce Research,** Talca, v. 9, n. 3, p.15-31, Sep. 2014. Disponível em: < http://www.scielo.cl/pdf/jtaer/v9n3/art03.pdf>. Acesso em: 18 abr. 2017.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI magazine**, v. 17, n. 3, p. 37-54, 1996. Disponível em: < http://www.csd.uwo.ca/faculty/ling/cs435/fayyad.pdf>. Acesso em: 06 mar. 2017.

G-8. Disponível em: <http://brasilescola.uol.com.br/geografia/g8.htm>. Acesso em: 22 abr. 2017.

GARCÍA, J. G. Gobierno abierto: transparencia, participación y colaboración en las administraciones públicas. **Innovar:** Revista de Ciencias Administrativas y Sociales, Bogotá, v. 24, n. 54, p.75-88, oct/dic. 2014. Disponível em: < http://www.redalyc.org/articulo.oa?id=81832222006>. Acesso em: 18 abr. 2017.

GIL, A. C. A pesquisa social. In: GIL, A. C. **Métodos e técnicas de pesquisa social.** 2. ed. São Paulo: Atlas, 1989. Cap. 3. p. 43-59.

GIL, A. C. O delineamento da pesquisa. In: GIL, A. C. **Métodos e técnicas de pesquisa social.** 2. ed. São Paulo: Atlas, 1989. Cap. 6. p. 70-80.

GIL, A. C. **Métodos e técnicas de pesquisa social**. 6. ed. São Paulo: Atlas, 2008.

INOMATA, D. O. **O Fluxo da informação tecnológica**: uma análise no processo de desenvolvimento de produtos biotecnológicos. 2012. 283 f. Dissertação (Mestrado) - Curso de Programa de Pós-graduação em Ciência da Informação, Universidade Federal de Santa Catarina, Florianópolis, 2012. Disponível em: <https://repositorio.ufsc.br/bitstream/handle/123456789/99498/305147.pdf?sequenc =1&isAllowed=y>. Acesso em: 02. Out. 2017.

FERREIRA, P. M. da S. **Aplicação de Algoritmos de Aprendizagem Automática para a Previsão de Cancro de Mama**. 2010. 207 p. Dissertação (Mestrado) - Engenharia de Redes e Sistemas Informáticos, Faculdade de Ciências da Universidade do Porto, Porto, 2010. Disponível em:< https://www.dcc.fc.up.pt/~ines/aulas/1516/DM1/TesePedroFerreira.pdf>. Acesso em 02. Out. 2017.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE. **Censo demográfico 2010:** Características gerais da população, religião e pessoas com deficiência. Rio de Janeiro, 2010. Disponível em: <http://biblioteca.ibge.gov.br/visualizacao/periodicos/94/cd_2010_religiao_deficiencia .pdf>. Acesso em: 22 abr. 2017.

JARDIM, J. M. A lei de acesso à informação pública: dimensões político-informacionais. **Tendências da Pesquisa Brasileira em Ciência da Informação**, Rio de Janeiro, v. 5, n. 1, p. 1-20, 2012. Disponível em: < http://inseer.ibict.br/ancib/index.php/tpbci/article/view/68/110 >. Acesso em: 18 abr. 2017.

LIMA, T. DOS S. **Estudo comparativo dos algoritmos de classificação da ferramenta *WEKA***. 2005. 76 f. Trabalho de Conclusão de Curso (Graduação) - Curso de Sistemas de Informação, Centro Universitário Luterano de Palmas, Palmas, 2005. Disponível em: < http://arquivo.ulbra-to.br/ensino/43020/artigos/relatorios2005-2/Arquivos/Alan%20O%20S%20-

%20Estagio%20Supervisionado%20em%20Sistemas%20de%20Informacao.pdf>. Acesso em: 02. Out. 2017.

MÁCHOVÁ, R.; LNěNIčKA, M. Evaluating the quality of open data portals on the national level. **Journal of Theoretical and Applied Electronic Commerce Research,** Talca, v. 12, n. 1, p.21-41, Jan. 2017. Disponível em: < http://www.scielo.cl/pdf/jtaer/v12n1/art03.pdf>. Acesso em: 18 abr. 2017.

MARCHIORI, Patricia Zeni. A ciência e a gestão da informação: compatibilidades no espaço profissional. **Ciência da Informação,** Brasília, v. 31, n. 2, p.72-79, ago. 2002. Disponível em: <http://www.scielo.br/pdf/ci/v31n2/12910.pdf/>. Acesso em: 10 nov. 2017.

MARTINEZ, Luís; CASAL, Ricardo; JANEIRO, João. **Sistemas de apoio à decisão clínica**. Porto: Faculdade de Medicina do Porto, 2009. 16 p.

MCGEE, J.; PRUSAK, L. **Gerenciamento estratégico da informação:** aumente a competitividade e a eficiência de sua empresa utilizando a informação como uma ferramenta estratégica. 6.ed. Rio de Janeiro: Campus, 1994. 244p.

MONTEZANO, N. S. **A importância da Gestão da Informação para as empresas e para a atuação do Secretário Executivo**: uma Revisão Bibliográfica. Disponível em: < http://www.secretariadoexecutivo.ufv.br/docs/Nuriane.pdf>. Acesso em: 21 Set. 2016.

OPEN KNOWLEDGE INTERNATIONAL. **Why open data?** Open data, especially open government data, is a tremendous resource that is as yet largely untapped. 2017. Disponível em: <https://okfn.org/opendata/why-open-data/>. Acesso em: 06 mar. 2017.

PORÉM, M. E.; SANTOS, V. C. B. dos; BELLUZZO, R.C.B. Vantagem competitiva nas empresas contemporâneas: a informação e a inteligência competitiva na tomada de decisões estratégicas**. Intexto**, Porto Alegre, n. 27, p.183-199, dez. 2012. Disponível em: Acesso em: 18 Nov. 2016.

PRASS, Fernando Sarturi. **ESTUDO COMPARATIVO ENTRE ALGORITMOS DE ANÁLISE DE AGRUPAMENTOS EM DATA MINING.** 2004. 71 f. Dissertação (Mestrado) - Curso de Mestrado em Ciência da Computação, Universidade Federal de Santa Catarina, Florianópolis, 2004. Disponível em: <http://fp2.com.br/blog/wp-content/uploads/2012/07/Dissertacao_Fernando_Prass_Data_Mining_Cluster_Analysis.pdf>. Acesso em: 18 Abr. 2017.

RAMPÃO, T. de S. **Mineração de dados em bases jurídicas:** um estudo de caso. 2016. 159 f. Trabalho de Conclusão de Curso (Graduação) - Curso de Gestão da Informação, Universidade Federal do Paraná, Curitiba, 2016. Disponível em: <http://www.decigi.ufpr.br/pesquisa-4/monografias>. Acesso em: 09 maio 2017.

SILVA, E. L. da; MENEZES, E. M. **Metodologia da pesquisa e elaboração de dissertação**. 4. ed. Florianópolis: UFSC, 2005. 138 p. Disponível em:

<https://projetos.inf.ufsc.br/arquivos/Metodologia_de_pesquisa_e_elaboracao_de_teses_e_dissertacoes_4ed.pdf>. Acesso em: 18 abr. 2017.

SILVA, M. P. dos S. **Mineração de Dados**: Conceitos, Aplicações e Experimentos com *WEKA*. Mossoró, 2004. Disponível em: <http://www.lbd.dcc.ufmg.br/colecoes/erirjes/2004/004.pdf>. Acesso em: 18 abr. 2017.

SOARES JUNIOR, J. S.; QUINTELLA, R. H. Descoberta de conhecimento em bases de dados públicas: uma proposta de estruturação metodológica. **Revista de Administração Pública,** Rio de Janeiro, v. 39, n. 5, p.1077-1107, out. 2005. Disponível em: <http://bibliotecadigital.fgv.br/ojs/index.php/rap/article/view/6580/5164>. Acesso em: 18 abr. 2017.

SOCZEK, F. C.; ORLOVSKI, R. **Mineração de Dados**: Conceitos e aplicação de algoritmos em uma Base de Dados na área da saúde. Guarapuava, v. 1, p. 1-25, mar. 2014. Disponível em: < http://semanaacademica.org.br/artigo/mineracao-de-dados-conceitos-e-aplicacao-de-algoritmos-em-uma-base-de-dados-na-area-da-saude >. Acesso em: 02 Out. 2017.

TARAPANOFF, K. Informação, conhecimento e inteligência em corporações: relações e complementaridade. In: TARAPANOFF, Kira (Org.). **Inteligência, Informação e Conhecimento.** Brasília: Unesco, 2006. p. 19-36. Disponível em: <http://unesdoc.unesco.org/images/0014/001469/146980por.pdf>. Acesso em: 27 set. 2017.

UNIVERSIDADE DE SÃO PAULO. **Declaração universal dos direitos humanos**. 1948. Disponível em: < http://www.direitoshumanos.usp.br/index.php/Declara%C3%A7%C3%A3o-Universal-dos-Direitos-Humanos/declaracao-universal-dos-direitos-humanos.html>. Acesso em: 18 abr. 2017

VALDIVIA, R.; NAVARRETE, M.; ARACENA, D. Oportunidades en datos abiertos. **Ingeniare**: Revista Chilena de Ingeniería, v. 22, n. 4, p.458-459, 2014. Disponível em: < http://www.scielo.cl/pdf/ingeniare/v22n4/art01.pdf>. Acesso em: 18 abr. 2017.

VENTURA, M. Lei de acesso à informação, privacidade e a pesquisa em saúde. **Perspectivas,** Rio de Janeiro, v. 29, n. 4, p.636-638, abr. 2013. Disponível em: < http://www.scielo.br/pdf/csp/v29n4/02.pdf>. Acesso em: 18 abr. 2017.

YOSHIDA, N. D. Análise bibliométrica: um estudo aplicado à previsão tecnológica. **Future Studies Research Journal,** São Paulo, v. 2, n. 1, p.52-84, jan/jun. 2010. Disponível em: < https://revistafuture.org/FSRJ/article/viewFile/45/68>. Acesso em: 18 abr. 2017.

ANEXO A - TRANFORMAÇÃO DOS ATRIBUTOS NUMÉRICOS PARA CATEGÓRICOS SEGUNDO VALORES DO CAGED

(Continua)

VARIÁVEIS	VALOR NA FONTE	TRANSFORMADO PARA
Faixa Empr Início Jan	1	ATE 4
	2	DE 5 A 9
	3	DE 10 A 19
	4	DE 20 A 49
	5	DE 50 A 99
	6	DE 100 A 249
	7	DE 250 A 499
	8	DE 500 A 999
	9	1000 OU MAIS
	Nulo	IGNORADO
Grau Instrução	1	Analfabeto
	2	Até 5ª Incompleto
	3	5ª Completo Fundamental
	4	6ª a 9ª Fundamental
	5	Fundamental Completo
	6	Médio Incompleto
	7	Médio Completo
	8	Superior Incompleto
	9	Superior Completo
	10	MESTRADO
	11	DOUTORADO
	Nulo	IGNORADO
Ind Aprendiz	1	SIM
	0	NÃO
Raça Cor	1	INDIGENA
	2	BRANCA
	4	PRETA
	6	AMARELA
	8	PARDA
	9	NAO IDENT
	Nulo	IGNORADO
Sexo	1	MASCULINO
	2	FEMININO
	Nulo	IGNORADO
Tipo Defic	1	FISICA
	2	AUDITIVA
	3	VISUAL
	4	Intelectual (Mental)

(Continuação)

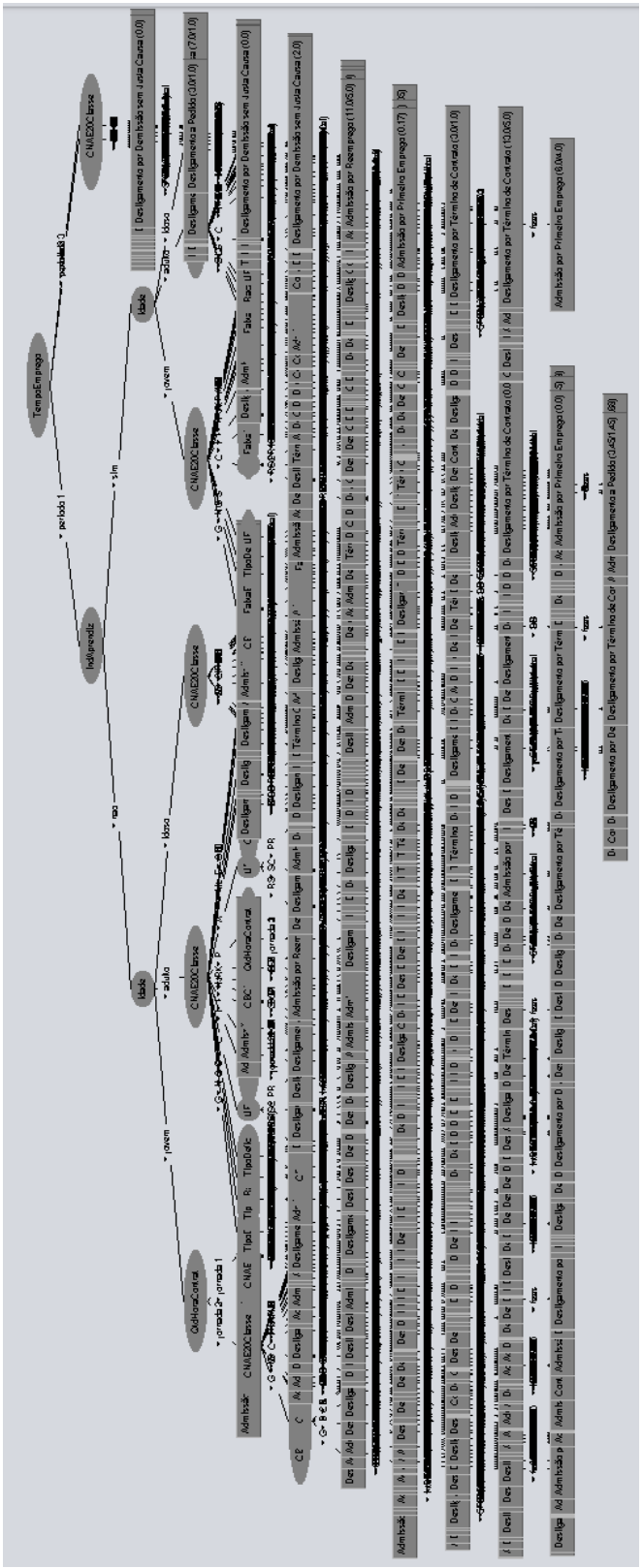
VARIÁVEIS	VALOR NA FONTE	TRANSFORMADO PARA
Tipo Defic	5	MULTIPLA
	6	REABILITADO
	Nulo	IGNORADO
Tipo Mov Desagregado	1	Admissão por Primeiro Emprego
	2	Admissão por Reemprego
	3	Admissão por Transferência
	4	Desligamento por Demissão sem Justa Causa
	5	Desligamento por Demissão com Justa Causa
	6	Desligamento a Pedido
	7	Desligamento por Aposentadoria
	8	Desligamento por Morte
	9	Desligamento por Transferência
	10	Admissão por Reintegração
	11	Desligamento por Término de Contrato
	25	Contrato Trabalho Prazo Determinado
	43	Término Contrato Trabalho Prazo Determinado
	-1	IGNORADO
UF	41	Paraná
	42	Santa Catarina
	43	Rio Grande do Sul

FONTE: ADAPTADO DO CAGED (2017).

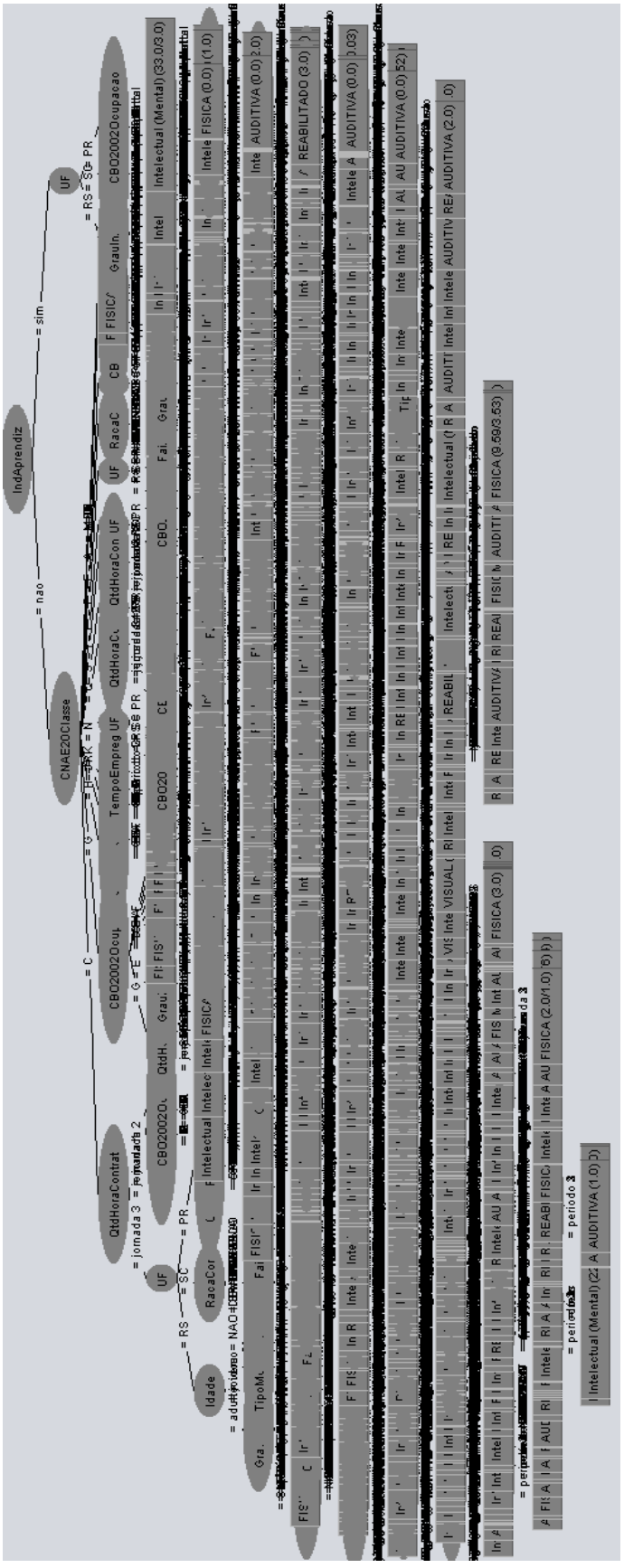
ANEXO B - INDICADORES DE AVALIAÇÃO DE DESEMPENHO DOS RESULTADOS DE ALGORITMOS DE CLASSIFICAÇÃO

```
Time taken to build model: 2209.39 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      130243               50.4806 %
Incorrectly Classified Instances    127763               49.5194 %
Kappa statistic                          0.2453
Mean absolute error                      0.2034
Root mean squared error                  0.3286
Relative absolute error                 84.7277 %
Root relative squared error             94.8318 %
Total Number of Instances           258006

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area  Class
                 0,304     0,127     0,416       0,304   0,351       0,199    0,675     0,392     AUDITIVA
                 0,785     0,546     0,530       0,785   0,632       0,248    0,684     0,625     FISICA
                 0,046     0,018     0,230       0,046   0,077       0,061    0,615     0,154     VISUAL
                 0,333     0,037     0,494       0,333   0,398       0,354    0,795     0,405     REABILITADO
                 0,034     0,001     0,301       0,034   0,062       0,098    0,637     0,052     MULTIPLA
                 0,453     0,046     0,563       0,453   0,502       0,448    0,820     0,515     Intelectual (Mental)
Weighted Avg.    0,505     0,280     0,470       0,505   0,464       0,249    0,701     0,480

=== Confusion Matrix ===

     a     b     c     d     e     f   <-- classified as
 18023 34322  1058  2459    57  3445 |     a = AUDITIVA
 13764 88910  2014  4020    94  4531 |     b = FISICA
  4133 18411  1228  1391    50  1428 |     c = VISUAL
  3159 12858   411  8469    23   508 |     d = REABILITADO
   539  2049    76   129   119   549 |     e = MULTIPLA
  3672 11350   545   665    53 13494 |     f = Intelectual (Mental)
```

ANEXO C - RESULTADO GRÁFICO DO J48 NO EXPERIMENTO A1 (PADRÃO NO MERCADO FORMAL DAS PCD NA REGIÃO SUL DO BRASIL)



ANEXO D - RESULTADO GRÁFICO DO J48 NO EXPERIMENTO B1 (PERFIL DAS PCD NA REGIÃO SUL DO BRASIL)



APÊNDICE A – DADOS ORIGINAIS DO CAGED

(Continua)

	VARIÁVEIS	DESCRIÇÃO DA VARIÁVEL	CATEGORIAS	VALOR NA FONTE	TIPO DO DADO
1	**Admitidos Desligados**	Admissão ou Desligamento.	ADMISSAO	1	CAT: BINÁRIO
			DESLIGAMENTO	2	
2	**Competência Declarada**	Competência (mês/ano) em que a movimentação foi declarada	AAAAMM EXEMPLO:200902	999999	NUM: DISCRETO
3	**Município**	Município de localização do estabelecimento	MUN<99.9999>	999999	NUM: DISCRETO
			IGNORADO	Nulo	
4	**Ano Declarado**	Ano em que a movimentação foi declarada	AAAA	9999	NUM: DISCRETO
			IGNORADO	Nulo	
5	**CBO 2002 Ocupação**	Classificação Brasileira de Ocupações, criada em 2002	CBO <999999>	999999	NUM: DISCRETO
			IGNORADO	Nulo	
6	**CNAE 1.0 Classe**	Classe de Atividade Econômica segundo a classificação CNAE/95, convertida a partir da Classe CNAE 20, disponível a partir de 01/2008	CLASSE <99999>	99999	NUM: DISCRETO
			NAO INFORM	00000	
			IGNORADO	Nulo	
7	**CNAE 2.0 Classe**	Classe de Atividade Econômica, segundo classificação CNAE - versão 2.0	CLASSE <99999>	99999	NUM: DISCRETO
			NAO INFORM	00000	
			IGNORADO	Nulo	
8	**CNAE 2.0 Subclasse**	Subclasse de Atividade Econômica, segundo classificação CNAE - versão 2.0 a partir de 2006	CLAS <9999999>	9999999	NUM: DISCRETO
			NAO INFORM	00000	
			IGNORADO	Nulo	
9	**Faixa Empr. Início Jan**	Tamanho do estabelecimento em janeiro do ano de referência	ATE 4	1	NUM: DISCRETO
			DE 5 A 9	2	
			DE 10 A 19	3	
			DE 20 A 49	4	
			DE 50 A 99	5	
			DE 100 A 249	6	
			DE 250 A 499	7	
			DE 500 A 999	8	
			1000 OU MAIS	9	
			IGNORADO	Nulo	
10	**Grau Instrução**	Grau de instrução ou escolaridade	Analfabeto	1	NUM: DISCRETO
			Até 5ª Incompleto	2	
			5ª Completo Fundamental	3	
			6ª a 9ª Fundamental	4	
			Fundamental Completo	5	

(Continuação)

	VARIÁVEIS	DESCRIÇÃO DA VARIÁVEL	CATEGORIAS	VALOR NA FONTE	TIPO DO DADO
10	**Grau Instrução**	Grau de instrução ou escolaridade	Médio Incompleto	6	
			Médio Completo	7	
			Superior Incompleto	8	
			Superior Completo	9	
			MESTRADO	10	
			DOUTORADO	11	
			IGNORADO	Nulo	NUM: DISCRETO
11	**Qtd Hora Contrat**	Quantidade de horas contratuais por semana	<99>	99	NUM: DISCRETO
12	**IBGE Subsetor**	Subsetor Econômico segundo IBGE	Extrativa mineral	1	NUM: DISCRETO
			Indústria de produtos minerais não metálicos	2	
			Indústria metalúrgica	3	
			Indústria mecânica	4	
			Indústria do material elétrico e de comunicações	5	
			Indústria do material de transporte	6	
			Indústria da madeira e do mobiliário	7	
			Indústria do papel, papelão, editorial e gráfica	8	
			Ind. da borracha, fumo, couros, peles, similares, ind. diversas	9	
			Ind. química de produtos farmacêuticos, veterinários, perfumaria	10	
			Indústria têxtil do vestuário e artefatos de tecidos	11	

(Continuação)

	VARIÁVEIS	DESCRIÇÃO DA VARIÁVEL	CATEGORIAS	VALOR NA FONTE	TIPO DO DADO
			Indústria de calçados	12	
			Indústria de produtos alimentícios, bebidas e álcool etílico	13	
			Serviços industriais de utilidade pública	14	
			Construção civil	15	
			Comércio varejista	16	
			Comércio atacadista	17	
			Instituições de crédito, seguros e capitalização	18	
			Com. e administração de imóveis, valores mobiliários, serv. Técnico	19	
			Transportes e comunicações	20	
			Serv. de alojamento, alimentação, reparação, manutenção, redação	21	
			Serviços médicos, odontológicos e veterinários	22	
			Ensino	23	
			Administração pública direta e autárquica	24	
			Agricultura, silvicultura, criação de animais, extrativismo vegetal	25	
			Ignorado	Nulo	
13	**Idade**	Idade do trabalhador (quando acumulada representa a soma das idades)	<999>	999	CAT: RAZÃO

(Continuação)

	VARIÁVEIS	DESCRIÇÃO DA VARIÁVEL	CATEGORIAS	VALOR NA FONTE	TIPO DO DADO
14	**Ind. Aprendiz**	Indicador de movimentação referente a contrato de aprendizagem	SIM	1	CAT: BINÁRIO
			NÃO	0	
15	**Ind. Portador Deic**	Indicador se o empregado/servidor de portador de deficiência habilitado ou beneficiário reabilitado	SIM	1	CAT: BINÁRIO
			NÃO	0	
16	**Raça Cor**	Raça e Cor do Trabalhador - disponível a partir de 01/2007	INDIGENA	1	NUM: DISCRETO
			BRANCA	2	
17	**Salário Mensal**	Salário mensal em moeda corrente	PRETA	4	CAT: RAZÃO
18	**Saldo Mov**	Saldo de movimentação	AMARELA	6	CAT: BINÁRIO
			PARDA	8	
19	**Sexo**	Sexo	MASCULINO	1	CAT: BINÁRIO
			FEMININO	2	
			IGNORADO	Nulo	
20	**Tempo Emprego**	Tempo de emprego do trabalhador (quando acumulada representa a soma dos meses)	<999v9>	999,9	NUM: CONTÍNUO
21	**Tipo Estab**	Tipo de estabelecimento	CNPJ	1	CAT: DISCRETO
			CEI	3	
			NAO IDENTIF	9	
			IGNORADO	-1	
22	**Tipo Defic**	Tipo de deficiência/Beneficiário habilitado	FISICA	1	NUM: DISCRETO
			AUDITIVA	2	
			VISUAL	3	
			Intelectual (Mental)	4	
			MULTIPLA	5	
			REABILITADO	6	
			NAO DEFIC	0	
			IGNORADO	Nulo	
23	**Tipo Mov Desagregado**	Tipo de movimento	Admissão por Primeiro Emprego	1	NUM: DISCRETO
			Admissão por Reemprego	2	
			Admissão por Transferência	3	
			Desligamento por Demissão sem Justa Causa	4	

(Continuação)

	VARIÁVEIS	DESCRIÇÃO DA VARIÁVEL	CATEGORIAS	VALOR NA FONTE	TIPO DO DADO
			Desligamento por Demissão com Justa Causa	5	
			Desligamento a Pedido	6	
			Desligamento por Aposentadoria	7	
			Desligamento por Morte	8	
			Desligamento por Transferência	9	
			Admissão por Reintegração	10	
			Desligamento por Término de Contrato	11	
			Contrato Trabalho Prazo Determinado	25	
			Término Contrato Trabalho Prazo Determinado	43	
			IGNORADO	-1	
24	**UF**	Município de localização do estabelecimento	Rondônia	11	NUM: DISCRETO
			Acre	12	
			Amazonas	13	
			Roraima	14	
			Para	15	
			Amapá	16	
			Tocantins	17	
			Maranhão	21	
			Piauí	22	
			Ceará	23	
			Rio Grande do Norte	24	
			Paraíba	25	
			Pernambuco	26	
			Alagoas	27	
			Sergipe	28	
			Bahia	29	
			Minas Gerais	31	
			Espírito Santo	32	
			Rio de Janeiro	33	

(Continuação)

	VARIÁVEIS	DESCRIÇÃO DA VARIÁVEL	CATEGORIAS	VALOR NA FONTE	TIPO DO DADO
			São Paulo	35	
			Paraná	41	
			Santa Catarina	42	
			Rio Grande do Sul	43	
			Mato Grosso do Sul	50	
			Mato Grosso	51	
			Goiás	52	
			Distrito Federal	**53**	
25	**Bairros SP**	Bairros do Município de São Paulo	Bairros SP	1 – 1878	
			Outros municípios de SP	9999	
			Fora do Estado de SP	0	
			Ignorado	Nulo	
26	**Bairros Fortaleza**	Bairros do município de Fortaleza	Bairros Fortaleza	1-71	
			Outros municípios de CE	99	
			Fora do Estado de CE	0	
			Ignorado	Nulo	
27	**Bairros RJ**	Bairros do município do Rio de Janeiro	Bairros RJ	1-1750	
			Outros municípios de RJ	9999	
			Fora do Estado de RJ	0	
			Ignorado	Nulo	
28	**Distritos SP**	Distritos do município de São Paulo	<9999>	9999	
29	**Regiões Adm DF**	Regiões Administrativas do Distrito Federal	<9999>	9999	
30	**Mesorregião**	Mesorregião	<9999>	9999	
31	**Microrregião**	Microrregião	<99999>	99999	
32	**Região Adm RJ**	Região Adm RJ	<99>	99	
33	**Região Adm SP**	Região Adm SP	<999>	999	
34	**Região Corede**	Região Corede	<99>	99	
35	**Região Corede 04**	Região Corede 04	<99>	99	
36	**Região Gov SP**	Região Gov SP	<999>	999	
37	**Região Senac PR**	Região Senac PR	<99>	99	
38	**Região Senai PR**	Região Senai PR	<99>	99	

(Continuação)

	VARIÁVEIS	DESCRIÇÃO DA VARIÁVEL	CATEGORIAS	VALOR NA FONTE	TIPO DO DADO
39	**Região Senai SP**	Região Senai SP	<99>	99	
40	**Sub-Região Senai PR**	Sub-região Senai PR	<99>	99	

FONTE: ADAPTADO DO CAGED (2017).